Structure and function of metal-independent CAZy family 6 glycosyltransferase from Bacteroides ovatus

Pham, Tram

Award date:
2014

Awarding institution:
University of Bath

Structure and function of metal-independent CAZy family 6 glycosyltransferase from *Bacteroides ovatus*

Tram Pham

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Biology and Biochemistry

June 2014

Acknowledgements

First of all, I would like to thank the Vietnamese government and the Biotechnology Center of Ho Chi Minh City, Vietnam for their financial support which offered me an opportunity to pursue a higher degree in the UK. I am also grateful to Professor Ravi K. Acharya for his supervision and support during my PhD study. I would like to thank Professor Keith Brew from Florida Atlantic University, USA for the consistent supply of highly pure protein which is the main material of this thesis. I would like to thank Dr. Nethaji Thiyagarajan who introduced me step by step to crystallography and everyone in Laboratory 0.34 for all their help and a friendly working environment. I would like to thank Charlotte Harrison and Dr. Ami Miller for their kind assistance in proofreading the English in my thesis. I would also like to thank University of Bath for financial support while writing the thesis. I would like to thank Diamond Light Source for access to MX stations I04 and I04-1 that were used for X-ray diffraction data collection for the work reported in this thesis.

I would like to send my thanks to the All Nations Group for all their encouragement and support, which helped me get over my homesickness and get along with my new life in Bath. To my wonderful friends, Ngoc, Thao, Duoi Uoi, Trinh, Phuong, Diu Dang, Elvis, Anh, Giannina, Marta, Sharon, Green and Hyeon; thanks to your company during my daily life which made my life here more comfortable and enjoyable. I specially thank Dave and Anne Sixmith who provided me not only the best place to live but also a family to live with during my writing stage. Thanks for their encouragement and being “patronising”, which kept me on track to finish my thesis.

Last but not least, I send all my love to my parents and thanks for their unconditional love, being my strength, and my inspiration. This thesis is dedicated to them.

Contents

Figures

Tables

Abbreviations

Å	Angstrom
AC	Affinity Chromatography
APS	Ammonium persulphate
Bis-Tris	Bis(2-hydroxyethyl)imino-tris(hydroxymethyl)methane
CAZy	Carbohydrate Active enZymes
CCP4	Collaborative Computational Project Number 4
Da	Dalton
DLS	Dynamic Light Scattering
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
ExPASy	Expert Protein Analysis System
FAL	2'-fucosyllactose
Gal	Galactose
GalNAc	α-N-acetylgalactosamine
GOL	Glycerol
GTA	Histo blood group A glycosyltransferase
GTB	Histo blood group B glycosyltransferase
GTs	Glycosyltransferases
HEPES	4-(2-hydroxyethyl)-1-piperazine ethanesulfonic acid
HIC	Hydrophobic Interaction Chromatography
KDa	Kilodalton
kpsi	Kilopound per square inch
LB	Luria Bertani media
LLG	Log Likelihood Gain
M	Molarity
mAu	mili-absorbance unit
MES	2-(n-morpholino)-ethanesulfonic acid
min	Minute(s)
mM	Milimolar
MS	Mass spectrometry

MW	Molecular weight
NAG	β-N-acetylgalactosamine
PDB	Protein Data Bank
PEG	Polyethylene glycol
pI	Isoelectric point
RFZ	Rotation function Z-score
RMSD (or r.m.s.d)	Root mean square deviation
RNA	Ribonucleic acid
rpm	Revolutions per minute
SDS	Sodium dodecyl (lauryl) sulphate
SEC	Size Exclusion Chromatography
S_N2	Nucleophilic substitution reaction of second order
S_Ni	Nucleophilic substitution internal reaction
S_Ni like	Internal return mechanism
TB	Terrific broth media
TEMED	N,N,N',N'-Tetramethylethylenediamine
TFZ	Translation function Z-score
Tris-HCl	Tris(hydroxymethyl)aminomethane hydrochloride
UDP	Uridine 5'-diphosphate
UDP-2F-gal	UDP-2'-fluorogalactose
UDP-Gal	UDP-Galactose
UDP-GalNAc	UDP-N-acetylgalactosamine
UDP-Glc	UDP-Glucose
V	Volt

Amino acid abbreviation		
Amino acid	3 letters	1 letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Materials and Methods

All chemicals used in this project were from Sigma (UK) unless otherwise stated.

The pictures were created using Coot (Emsley and Cowtan, 2004), Procheck (Laskowski *et al.*, 1993), and Pymol (DeLano, 2010).

Abstract

Glycosyltransferases (GTs) are an enzyme superfamily responsible for the synthesis of glyconjugates by transferring the sugar moiety from an active donor to a specific acceptor, usually a protein or a polysaccharide. The glyconjugates, such as glycolipids and glycoproteins, play vital roles in physical maintenance of tissue structures, immune recognition and other biological activities. Understanding their catalytic mechanisms is therefore critical. There are two kinds of reaction defined based on the stereochemistry of the product carbohydrate moiety compared to the donor: retaining and inverting. Whilst the mechanism of inverting GTs is well-known, the catalytic process of retaining GTs is as yet unclear.

Glycosyltransferase family 6 (GT6), according to the Carbohydrate Active Enzyme (CAZy) database, is a retaining GT family which catalyses the transfer of α-galactose (α-Gal) or α-N-acetyl-galactosamine (α-GalNAc) to the 3-OH group of a β-linked Gal or GalNAc in an acceptor substrate. Most GT6s from vertebrates require a metal ion for their activity. The metal ion-dependence is linked to the AspXaaAsp (DXD) motif which is conserved among these enzymes. However, analysing sequences of GT6s from bacteria showed that the DXD motif was substituted by an AsnXaaAsn (NXN) sequence. One of two CAZy family 6 glycosyltransferases, BoGT6a from *Bacteroides ovatus*, which catalyses the transfer of GalNAc from UDP-GalNAc to the saccharide acceptor and UDP-GalNAc hydrolysis, was kinetically and structurally studied. This enzyme is fully active in the absence of metal ions. The structure of BoGT6a is strikingly similar to its mammalian homologues such as GTA, GTB and α-1,3-galactosyltransferase, but it has a shorter N-terminal region and a NXN motif instead of a DXD motif. This suggests that the substitution of the DXD motif with the NXN may affect the catalytic mechanism of the enzyme.

The structure of the enzyme in complex with its acceptor molecule 2'-fucosyllactose was obtained at 3.0 Å. Comparison of the X-ray crystallographic structures of BoGT6a in its native and acceptor bound forms demonstrated the conformational changes of the enzyme associated with acceptor binding. It also elucidated the impact of acceptor binding on enzyme conformation and the structural relationship between the enzyme and its homologues.

Structural snapshots of the BoGT6a Glu192Gln (E192Q) mutant processing its donor UDP-GalNAc were also obtained. The interactions between the enzyme and the donor provide an insight into the mechanistic role of the NXN motif and nearby amino acid residues in BoGT6a's metal-independent activity. Moreover, the high flexibility of the enzyme conformation when it interacts with the ligands provides a general picture of how the enzyme processes UDP-GalNAc.

Together, these structures illustrate how a significant divergence in catalytic properties can be accommodated by minor structural adjustments, and propose a role for the NXN motif, which replaces the DXD motif in the metal independent glycosyltransferases.

Publications

A significant part of the content of this thesis was published as follows:

(1) Most of the content of Chapter 2 appeared in *Nature Scientific Reports* [Thiyagarajan, N., Pham, T. T. K, Stinson, B., Sundriyal, A., Tumbale, P., Lizotte-Waniewski, M., Brew, K., and Acharya, K. R. (2012) Structure of a metal-independent bacterial glycosyltransferase that catalyzes the synthesis of histo-blood group A antigen. *Sci. Rep.* **2**, 940.]

(2) Most of the content of Chapter 3 appeared in *Journal of Biological Chemistry* [Pham, T. T. K., Stinson, B., Thiyagarajan, N., Lizotte-Waniewski, M., Brew, K., and Acharya, K. R. (2014) Structures of complexes of a metal-independent glycosyltransferase GT6 from *Bacteroides ovatus* with UDP-N-Acetylgalactosamine (UDP-GalNAc) and its hydrolysis products. *J. Biol. Chem.* **289**, 8041-8050.]

CHAPTER I

Introduction

1 Introduction

1.1 Protein crystallography

Structural biology has developed dramatically since the first protein structure was solved at 6 Å resolution (Kendrew *et al.*, 1958). There are now a wide range of techniques such as X-ray crystallography, Nuclear magnetic resonance (NMR) spectroscopy, and electron microscopy (EM) available to scientists. The knowledge of accurate molecular structures enriches our understanding of the fundamentals of biochemistry, such as reaction mechanisms. It is also a prerequisite for structure based functional studies which assist in the development of effective therapeutic agents and drugs.

X-ray crystallography is the most conventional and common technique in determination of protein structures. More than 89 % of structures deposited in the protein data bank (PDB) (Berman *et al.*, 2000) have been solved using this technique (May 2014).

The wavelength of X-rays, 1-2 Å, is comparable to the carbon-carbon (C-C) bond length, 1.5 Å, making them a powerful tool for providing atomic information about protein structures. X-rays are diffracted off the many identical and regularly ordered protein molecules in a crystal, and the X-ray diffraction data is processed by computer to generate maps of electron density. Crystallographers can interpret the protein structure from these. A drawback of crystallography is that crystallisation is a difficult and time-consuming process. Nonetheless, this technique has become more powerful due to both improved software and hardware.

A standard crystallography process begins with protein production, including expression and purification. Once a sufficient quantity of high quality protein has been obtained, crystallisation can be performed either manually or automatically by usage of robots. This is the most crucial step because without a crystal there is, of course, no crystallography. A good quality crystal will give high resolution diffraction data, using either an in-house X-ray source or the synchrotron source. The diffraction data can then be processed by various specialised programs to derive the structure of the protein. An outline of this process is given in Figure 1.

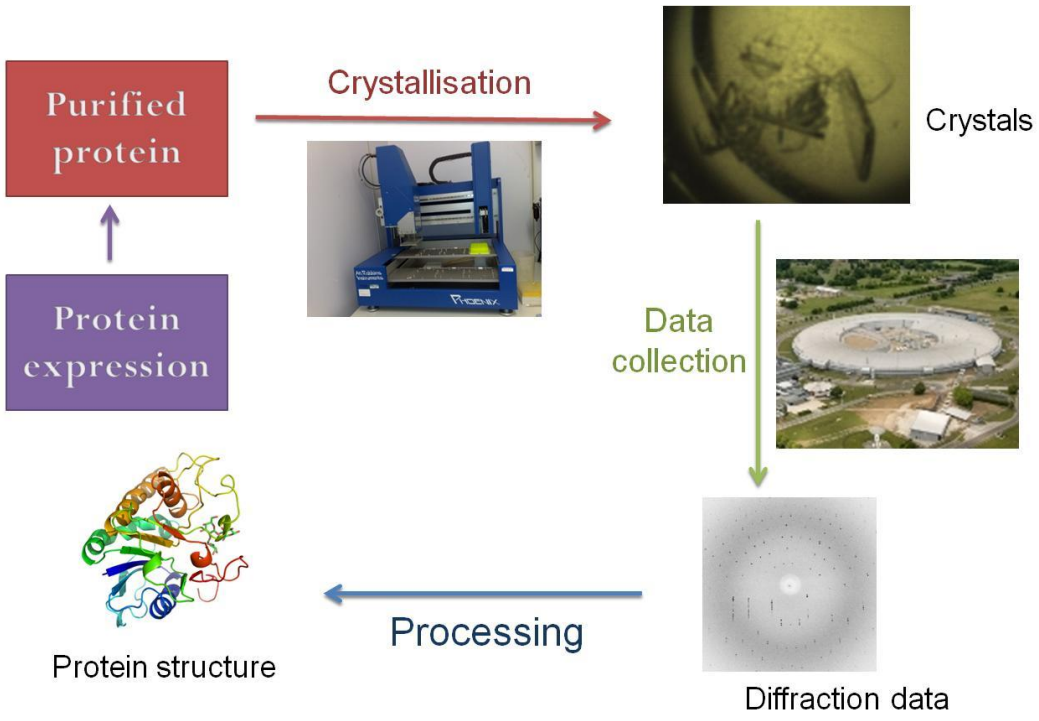


Figure 1. A general process of protein structure determination by using crystallography. The protein we are interested in is expressed in its required form with a high yield. It is then purified to be sufficiently homogenous and pure for crystallisation. Crystallisation may be performed robotically or manually. Diffraction data of the target crystals may be collected using an in-house X-ray source, or the synchrotron source. The final result, which is the structure of the target protein, is achieved by data processing on computers.

1.1.1 Preparation of protein

The first step of the protein structure determination process is producing an adequate supply of pure protein (Figure 1). Limited protein sources, which were mainly isolated directly from the original hosts, proved a barrier to crystallisation; this was particularly challenging for human proteins. However, once recombinant DNA technology had been applied to protein production, it became significantly easier to produce the necessary quantities of pure protein. The protein structure database has grown rapidly, and there were 100326 available structures in PDB at the time when the thesis was written (May, 2014).

Applications of various expression systems enable researchers to obtain their target proteins either from prokaryotes or eukaryotes. Each system has its own advantages

and disadvantages, which are summarised in table 1. *Escherichia coli* still remains the most popular expression system used in crystallography because of their low-cost, short life cycle, high yield and ease of genetic manipulations.

A commonly used *E. coli* strain in laboratories is BL21 (λDE3) Codon plus RIL/RP. This strain contains the T7-RNA polymerase promoter for faster protein translation, disabled *lon* and *omp*T proteases for new protein protection, and *arg*U, *ile*Y, *leu*W and *pro*L tRNA genes which recognise some rare codons such as AGG, AGA, AUA, CUA and CCC (Terpe, 2006, Kane, 1995). Proteins expressed in *E. coli* are occasionally in inclusion bodies, which are frequently incorrectly folded and aggregated. Hence insolubly expressed proteins first need to be solubilised, refolded, purified, concentrated and buffer exchanged to be homogeneous, pure, soluble, and stable at a high concentration before the crystallisation process.

Many purification techniques can be applied depending on the target protein's characteristics, such as affinity chromatography, size exclusion chromatography (or gel filtration), ion exchange chromatography, and hydrophobic interaction chromatography.

Affinity chromatography (AC) is the most varied and dominant chromatographic method for purification of a specific protein. It is based on highly specific biological interactions between two molecules, such as interactions between enzyme and substrate (such as between glutathione S-transferase and glutathione), receptor and ligand (such as between histidine-tagged proteins and metal ions), or antibody and antigen (such as antibodies binding to protein A or protein G). In a typical AC process, the sample is applied under conditions that favour specific binding to the ligand. Elution is performed specifically, using a competitive ligand, or non-specifically, by changing the pH, ionic strength, or polarity. The high selectivity of AC enables many separations to be achieved with high purity in a single step. When higher purity is required, one or more additional purification steps may be required, such as size exclusion chromatography.

Table 1. Advantages and drawbacks of protein expression systems (Sodoyer, 2004)

System	Advantages	Drawbacks
Prokaryotic		
Escherichia coli	High yield Large choice of genetic elements Low cost	No post-translational modifications
Bacillus	Secretion Low protease Low cost	No post-translational modifications
Eukaryotic		
Mammalian cells	Secretion Suitable for complex molecules	Additives Low yield
Insect cells	High yield Simple media Viral safety	Glycosylation profile
Vegetal	Biomass Secretion Viral safety	Glycosylation profile
Yeast	Biomass Secretion	Glycosylation profile
Nonconventional yeast	Growing capacity in extreme conditions/waste material	Genetics still need to be explored
Trypanosome	Mammalian-like glycosylation	Genetics still need to be explored
Transgenic animals	Suitable for complex molecules	Time consuming Restricted to very high added-value products
Cell-free translation		
CECF (continuous-exchange cell-free) or CFCF (continuous-flow cell-free) systems	Suitable for toxic molecules Incorporation of unnatural amino acids	Glycosylation Disulphide bond formation

Size exclusion chromatography (SEC) is usually the last step in the purification process. It improves the purification result by separating protein samples based on molecular size. A crude protein sample is applied to a column of porous resin. Proteins of different size pass through the column at different rates. Larger molecules take a shorter route through the column and thus are released from the column earlier than smaller molecules which penetrate into the porous resin and so have a longer retention time. The resolution of SEC depends on the size of the pores in the resins and the length of the column. The resolution of SEC is greater with finer resins and longer columns.

Generally, purification is a multistep process. Depending on the characteristics of the target protein and the impurities present, one or more of the methods listed above are carefully selected. These are then used in different combinations to maximise the purity of the target protein. The purity of the protein can be analysed using protein gel electrophoresis or modern techniques such as Dynamic Light Scattering and Mass Spectrometry.

Dynamic Light Scattering (DLS) is a spectroscopic technique generally used to determine the size distribution of protein molecules in solution. It is also a good indicator for any contaminants or heterogeneity of the protein sample. Mass spectrometry (MS) is another powerful tool in protein crystallisation. It can be used for analysing the expression of recombinant protein, assessing the purity of a preparation, or checking for heavy atom derivatives (during the structure determination process). It also gives information on the nature of a protein construct, for example the mass, whether it is in complex with its substrate(s) or in its apo form, whether it is monomeric or oligomeric. (Cohen, 1996, Carte *et al.*, 2000).

When the final protein is homogeneous, active, pure and stable in a suitable buffer at the required concentration (5-20 mg/ml), it will be ready for crystallisation trials.

1.1.2 Crystallisation

The principles of crystal growth have been intensely investigated for many years and the theoretical and practical aspects of crystallisation of molecules such as salts or small organic compounds are well established nowadays. However, although the

crystallisation of macromolecules such as protein, DNA and RNA is proving to be theoretically similar to small molecules, much work remains mainly based on trial experiments (McPherson, 1999). Three major stages of crystal formation are common to all systems; nucleation, growth, and cessation of growth.

A classical explanation of crystal nuclei formation and growth can be illustrated by a phase diagram (Figure 2). The solubility curve divides the concentration space into two areas - the undersaturated and supersaturated zones (Russo Krauss *et al.*, 2013). Each point on this curve corresponds to a concentration at which the solution is in equilibrium with the precipitating agent. In the soluble area, the solution is undersaturated and crystallisation will never take place. Above the solubility curve line is the supersaturation zone which is divided into three regions, namely metastable zone, labile zone and precipitation zone, based on the level of protein supersaturation.

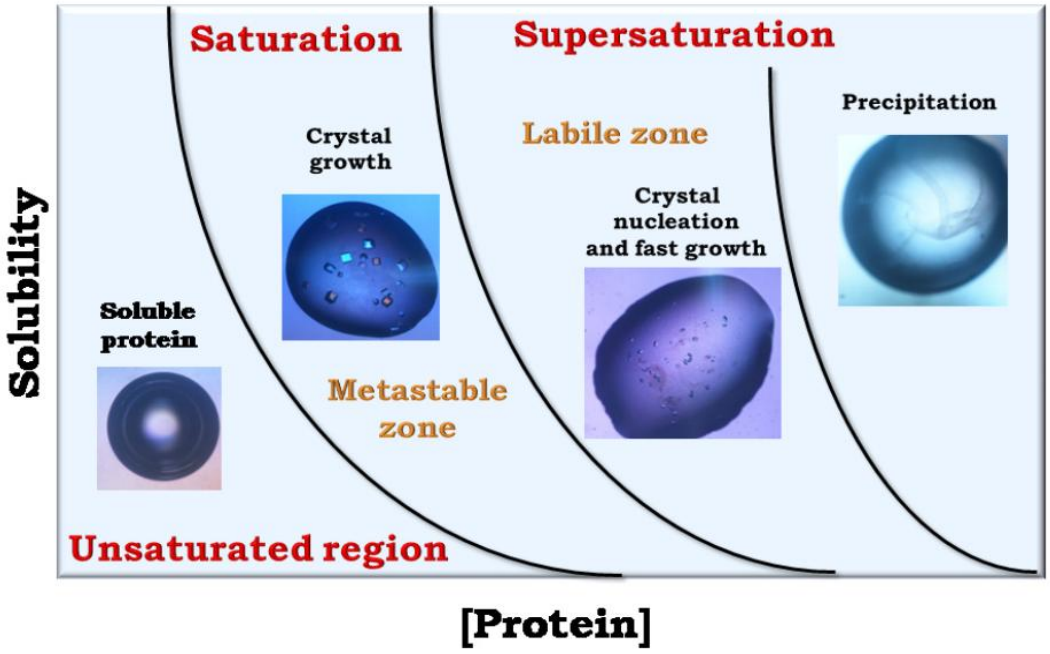


Figure 2. Crystallisation phase diagram. Schematic representation of a two-dimensional phase diagram illustrating the change of protein concentration against precipitating agent concentration. The concentration space is divided by the solubility curve into three areas corresponding to the unsaturated state, the saturated state (or metastable zone) and the supersaturated state of a protein solution. The supersaturated area comprises the labile zone and the precipitation zone (Russo Krauss *et al.*, 2013).

Crystal nuclei form and growth occurs in the labile zone in which protein concentration is just at intermediate supersaturation. The precipitation zone is where the protein concentration is so high that molecules immediately separate from the solution to form amorphous aggregates. The region of the labile zone near to the precipitation zone is where nucleus formation happens too fast, causing the formation of many microcrystals, sometimes they can be confused with precipitate. The last zone, the metastable zone, is where crystal growth is supported but nucleus formation cannot occur.

Based on this diagram, the general strategy of crystallisation is bringing the protein gradually from the soluble zone to the labile zone in order to obtain a single nucleus or a few nuclei. During nucleus formation, the solution will return to the metastable region. No more nuclei occur and the existing ones grow to bigger crystals at a decreasing rate, helping to avoid defect formation, until equilibrium is reached. However, in practice, it is difficult to identify these ideal conditions because each protein behaves differently in different solutions and a single experimental parameter change can concurrently influence several aspects of a crystallisation experiment.

Modern techniques have provided a clearer understanding of the crystallisation process. For example, atomic force microscopy (AFM) has proved a powerful tool in observing the general phenomenon of crystal growth (McPherson *et al.*, 2003). However, these high-tech methods are expensive and only available in a few crystallography laboratories. Crystallisation is often considered the main hurdle of protein structure determination. Obtaining suitable single crystals is the least understood step in the X-ray structural analysis of a protein. Protein crystallisation is mainly a trial-and-error procedure in which the most general crystallisation strategy is to bring the protein to a point only slightly above its saturation point as slowly as possible.

1.1.2.1 Crystallisation techniques

There are many crystallisation techniques that have been developed based on the principle that protein solubility decreases gradually until reaching a solid phase (crystal). Conventional and popular methods that have been employed by

crystallographers include batch, dialysis, vapour diffusion and free interface diffusion (FID). Although these methods differ, they all share the overarching aim of bringing the protein to the nucleation and metastable zone (Figure 3). Recently the gel crystallisation technique, already successfully applied to small molecules, has been developed for macromolecule crystallisation. The counter-diffusion in gel approach, introduced by Garcia-Ruiz (2003), appears particularly powerful as a method to produce many high quality crystals suitable for X-ray diffraction work.

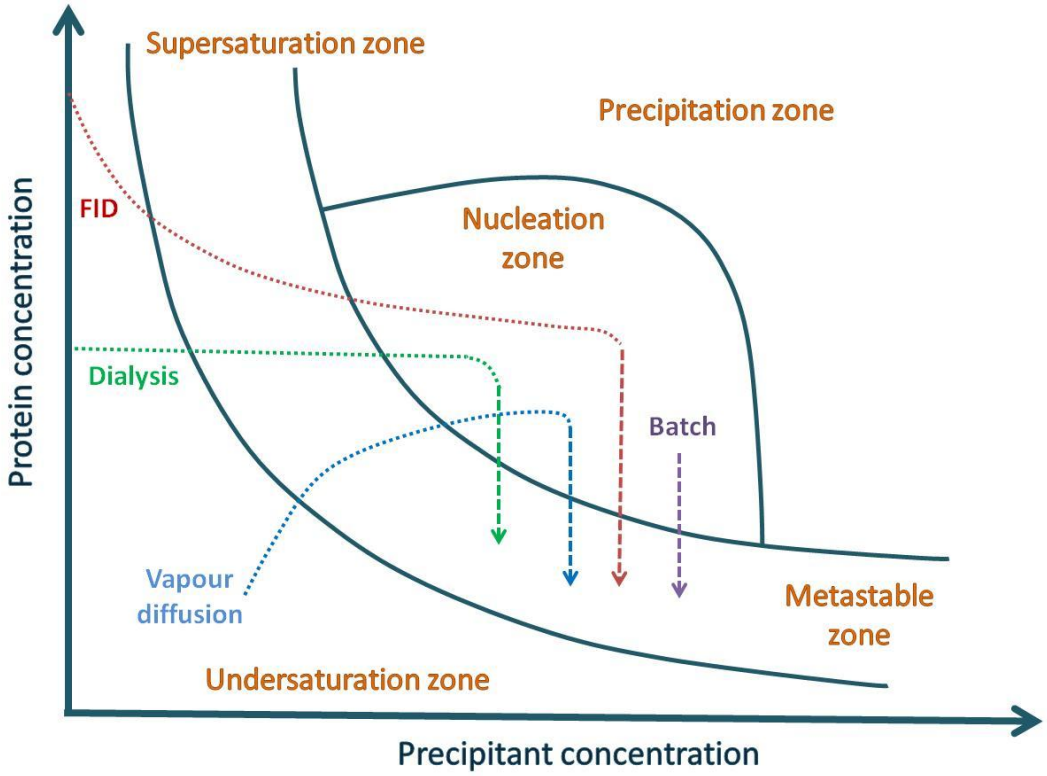


Figure 3. A simplified protein crystallisation phase diagram. The different routes of reaching nucleation and metastable zones for the four main crystallisation techniques (picture adapted from the review of Russo Krauss *et al.* (2013)).

The most popular technique in crystallisation is vapour diffusion in which the protein is brought into the nucleation zone by slowly increasing both protein concentration and precipitant concentrations simultaneously (Figure 3). The increase is caused by water evaporation and diffusion from the droplet containing a mixture of protein and reservoir solution into the reservoir solution. This technique is widely used because it is easy to manipulate conditions and apply high-throughput

screening, as well as optimisation of crystallisation conditions (Russo Krauss *et al.*, 2013). The technique consists of the hanging drop method, the sitting drop method and the sandwich drop method. Each method has different advantages and disadvantages, which are discussed below.

The hanging drop method is a simple approach in which a 2-6 µl droplet of protein and reservoir solution is placed on a cover slip. The cover slip is inverted such that the drop is facing the reservoir solution in a well and the well is sealed with grease (Figure 4A). The equilibrium rate in this method is highest among three methods because of the large surface area of the drop exposed to the reservoir solution. This method is mainly used in the optimisation stage of crystallisation procedure. Its disadvantages are difficulty in automation and incapacity in working with large volumes of sample. These drawbacks are overcome by the other methods.

The sitting drop method, in which the mixture of protein and reservoir solution is placed on a small bridge inside the well containing reservoir solution, can be used with a large volume of sample (Figure 4B). The biggest advantage of this method is that the whole procedure can be performed automatically by a robot, which is popular in all structural biology laboratories for crystallisation screening.

The last method in this technique, the sandwich drop method, is less popular than the other two, but in some cases, when crystallographers need to slow down the equilibration process, this method is a good choice. Because the droplet is placed between two cover slips (Figure 4C), the surface area of the drop is reduced, leading to a slower water vapour diffusion rate from the drop to the reservoir solution and as a consequence, better protein crystals can sometimes be produced.

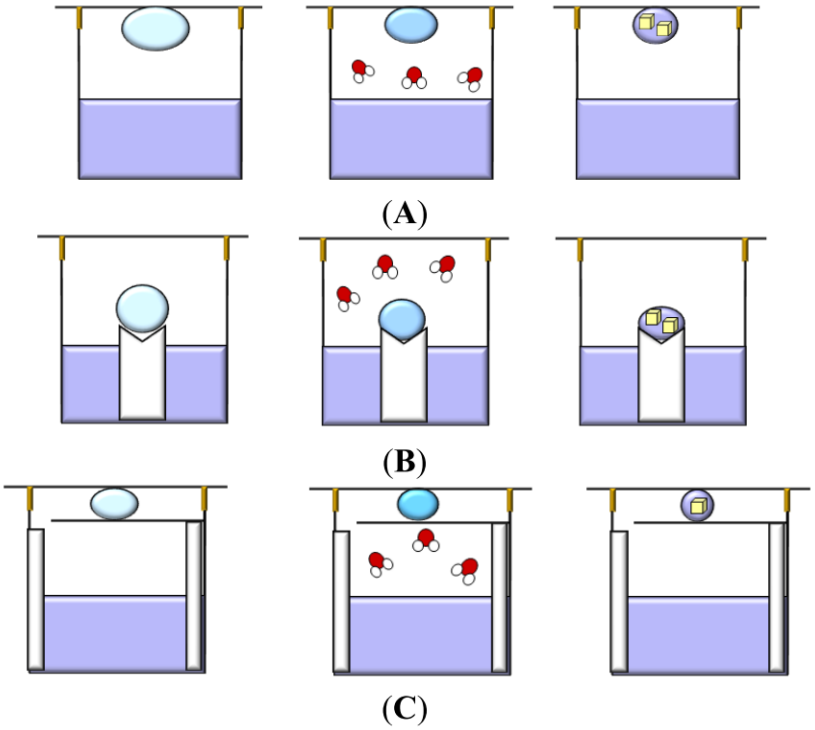


Figure 4. Vapour-diffusion technique. (A) hanging drop method, (B) sitting drop method, and (C) sandwich drop method (Russo Krauss *et al.*, 2013).

1.1.2.2 Screening and optimizing crystallisation

To obtain protein crystals, crystallographers generally go through two steps: (i) screening crystallisation conditions to identify which chemical, biochemical and physical conditions can produce some crystalline material, which is sometimes insufficient for diffraction data collection, and (ii) optimizing crystallisation based on the "hits" obtained from previous steps to achieve superior crystals which are suitable for diffraction data collection.

1.1.2.2.1 Screening

Protein crystallisation mainly relies on screening techniques with assistance from robots. Most of the robots commercially available nowadays are very practical for crystallisation condition screening purpose because they can handle nanolitre scale volumes of samples and eliminate human errors in repeating hundreds of experiments. Typical experiments when working with a robot involve setting up 96-well plates, enabling the analysis of 96 conditions simultaneously. To set up a 96

well plate only ~20 µl of protein, depending on dropsize, is required and the protein concentration usually ranges from 1-20 mg/ml.

The choice of which solution conditions to screen has been greatly simplified over the years. One source of solution conditions is The Biological Macromolecular Crystallisation Database, which is a free online collection of published successful crystallisation conditions for biological macromolecules (http://wwwbmcd.nist.gov:8080/bmcd/bmcd.html) (Gilliland *et al.*, 1994). In practice, however, one of the many commercial screening kits is used, based on the sparse matrix approach first proposed by Jancarik & Kim (Jancarik and Kim, 1991). These cover most of the efficient precipitants and pH, as well as a variety of other useful solution conditions, and can be purchased in ready-to-use form (for example, see http://www.hamptonresearch.com).

If crystals are not produced, the screens are usually moved to a cold room and allowed to re-equilibrate. Ultimately, if crystals still do not appear, the purity of the protein is often the first thing to be considered. Purification steps need to be improved to get a higher level of purity. The buffer used during protein purification can also be considered for protein activity, solubility and stability. One can also try to improve protein surface interactions by mutating one or more residues, which may make crystallisation more favourable. For example, surface lysine methylation (SLM), in which lysine residues are chemically methylated, can be used to improve the probability of protein crystallisation (Sledz *et al.*, 2010).

If crystals are produced during the screening trials, they can be used directly for X-ray analysis if their quality meets requirements for diffraction data collection, or further optimisation is conducted to improve their size and diffraction quality.

1.1.2.2.2 Optimisation

Optimisation involves varying the chemical and physical parameters around those of the reagent mixture that yielded your crystals and searching crystallisation parameter space by small increments away from the starting point.

During nucleation, if a high level of supersaturation occurs, then nuclei formation happens quickly. This leads to crystals of poor quality, such as being too small, of high mosaicity or having clusters of needles. In such cases, either reducing the level of supersaturation or seeding a metastable, supersaturated protein solution with crystals from earlier trials can be tried. There are two general seeding methods, namely micro seeding in which microcrystals are used as seeds, and macro seeding, in which a single crystal of a size 5-50 µm is used as a seed (Bergfors, 2003).

One disadvantage of seeding with microcrystals is that the number of seeds is uncontrollable. It is possible that too many nuclei will be introduced into the fresh supersaturated solution and plenty of crystals may appear, but none of them may be suitable for X-ray diffraction analysis. In practice, microcrystals from a parent solution are serially diluted between 10-2 and 10-7 to obtain the optimum amount of nuclei (Bergfors, 2003). Once seeds are ready, micro seeds can be introduced into the new drop in many different ways, such as using a seeding wand which is dipped into the microcrystal diluted solution to take seeds and then touched, stirred, or streaked across the surface of the new drop.

In macro seeding, crystals used as seeds should be large enough to be manipulated and transferred under a microscope. Like micro seeding, controlling the number of seeds is still a problem in this technique because there may be microcrystals adhering to the surface of the seed crystal. To avoid this, the macro seeds should be washed thoroughly by passing them through a sequence of intermediate transfer solutions. The washing step not only removes microcrystals but also improves the surface of the seed crystals and as consequence, new growth patterns may be induced when they are transferred to new protein solution (McPherson and Gavira, 2014). One important thing that should be considered is that the new solution must be supersaturated with respect to protein, but not extremely so, in order to ensure slow and ordered growth. It is usually recommended to start with the same condition of reservoir solution but lower protein concentration because protein concentration is linearly related to its solubility (Bergfors, 2003).

Seeding is often a useful technique for initiating the growth of crystals or inducing nucleation and growth at a lower level of supersaturation than might otherwise

spontaneously occur. This technique is also used with poor crystals of the target protein or even with oils and precipitates (Bergfors, 2003, Gavira *et al.*, 2011) . In some cases, seeds can be heterogeneous or epitaxial nucleants, such as fibres, animal hairs, epoxide coatings, and nanoscale etched surfaces of graphite and silicon (McPherson, 1999).

The final step, prior to the X-ray data collection, is to know how to manipulate or to prepare these crystals for a proper and successful data collection, either at a synchrotron or using an in-house X-ray source.

1.1.2.3 Cryo protection and crystal mounting

The traditional method of mounting crystals involves direct mounting in a fine glass capillary, which contains a droplet of the reservoir solution along with the crystal, onto the goniometer head. The advantage of this method is that the orientations of crystals are controlled and crystals do not dry out. However, the amount of mother liquor in the capillary can adversely affect crystal diffraction and handling the glass capillary need to be done with care and requires practice. Popular nowadays, due to its ease of manipulation, is the small loop which is large enough to maintain a thin layer of mother liquor around the crystal. With various sizes (50-200nm) and types (round loops or mesh loops) available, it is easy to find a loop suitable for the target protein crystal.

Through many experiments, crystallographers found that very low temperatures can reduce radiation damage of crystals during the data collection process. Liquid nitrogen has been a good source for producing such low temperatures (100 Kelvin) for collection of crystal diffraction data. When performing experiments at this temperature, the formation of ice around the cooled crystal can reduce the order of crystal blocks of the crystal, increasing its moisacity. This problem, which can be identified as a ice-ring on the diffraction image at around 3.9 Å, reduces the quality of crystal diffraction data. Thus crystals are frequently flash cooled in the presence of cryoprotectants such as glycerol, PEG or even the mother liquor itself when it contains a high precipitant concentration. Searching for a good cryoprotectant is an empirical task, since each protein reacts differently to different cryoprotectants.

1.1.3 Diffraction data collection

Diffraction data collection is the process of capturing the scattering of X-rays by the electrons in the molecules constituting the target crystal. The X-ray beams come from an X-ray generator, at the heart of which is an X-ray tube, rotating anode or particle storage ring.

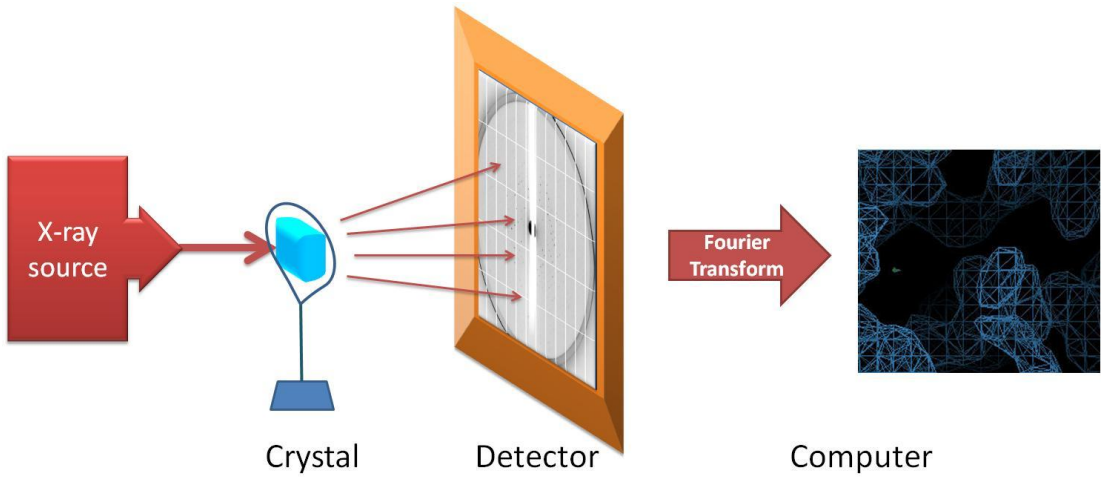


Figure 5. Diagram of a diffraction data collection process.

X-ray generators use a high voltage to accelerate electrons released by a hot cathode to a high velocity. These high velocity electrons collide with a metal target, the anode. Collisions with these incoming electrons displace electrons in the target material from lower to higher level orbitals. An X-ray photon is created when these electrons spontaneously return to the stable lower orbital.

X-ray tubes are composed of a water-cooled anode made of the target metal, whereas rotating anode tubes use a water-cooled rapidly rotating metal disk as an anode. Rotating anodes can create X-rays more than ten times as powerful as tubes with fixed anodes because they increase the metal surface exposed to the powerful electron bombardment from the cathode, reducing heat dissipation which is not as effectively reduced by water in the X-ray tubes. The particle storage rings are however the most powerful X-ray source, in which electrons circulate at velocities near the speed of light, driven by energy from radio-frequency transmitters and maintained in circular motion by powerful magnets. A charged body like an electron emits energy when forced into curved motion, and in accelerators, the energy is

emitted as X-rays. The intensity of X-rays is increased by wigglers which cause additional bending of the beam. The synchrotron source is not only powerful, which reduces data collection time, but also convenient due to the focusing mirror systems and monochromators which can produce powerful monochromatic X-ray at selectable wavelengths.

In a typical data collection experiment, a crystal is positioned in a beam of monochromatic X-ray radiation and the scattered rays are collected by a detector (Figure 5). X-rays passing through the crystal will cause the electrons of the molecules to oscillate. These oscillating charges then emit X-ray radiation of the same wavelength in all directions. Only scattered rays obeying Bragg's law that constructively interfere with each other can produce a diffraction pattern on a detector.

According to Bragg's law, an X-ray that reflects from the surface of a substance has travelled less distance than an X-ray which reflects from a plane of atoms inside the crystal. The penetrating X-ray travels down to the internal layer, reflects, and travels back over the same distance before being back at the surface. The distance travelled depends on the separation of the layers d and the angle θ at which the X-ray enters the material (Figure 6).

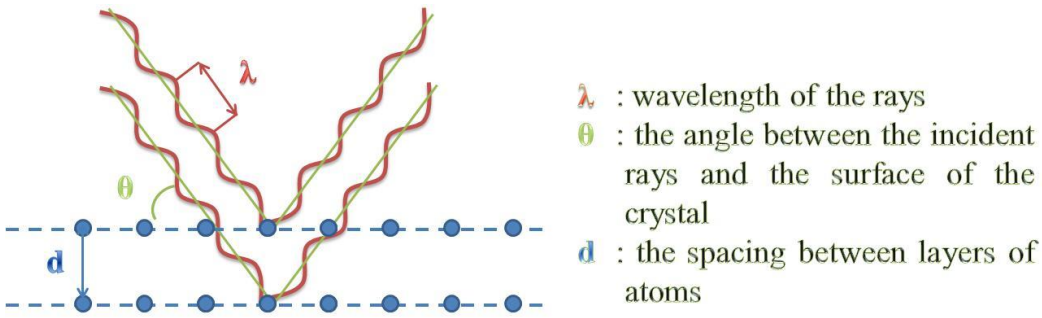


Figure 6. Diagram illustrates Bragg's law.

For this wave to be in phase with the wave which reflected from the surface it needs to have travelled a whole number of wavelengths while inside the material, which can be expressed in an equation

$$d = n\frac{1}{2}\left(\frac{\lambda}{\sin\theta}\right) \quad \text{(eq. 1)}$$

When n is an integer (1, 2, 3 etc.) the reflected waves from different layers are perfectly in phase with each other and produce a reflection on diffraction data. In other words, the waves that are not in phase will interfere destructively.

According to Bragg's law (eq. 1), the diffracted X-rays are only affected by the crystal structure or in other words, the geometry (size and shape) of the crystal unit cell, and the wavelength of the X-rays.

Firstly, the shape and the symmetry of the unit cell define the directions of the diffracted beams. The larger and higher the symmetry of the unit cell, the more diffracted beams or reflections can be observed. Secondly, the effectiveness of interference of the diffracted rays in each direction, and therefore the intensity of each diffracted ray, depends on the arrangement of all the atoms within the unit cell. In other words, the intensity of each individual reflection depends on the positions of all atoms in the unit cell. It is, therefore, not possible to solve only a selected, small part of the crystal structure without modelling the rest of it, in contrast to other structural techniques such as NMR or extended X-ray absorption fine structure spectroscopy which can describe only part of the molecule.

The entirety of a crystal can be described by its unit cell. Because a crystal is composed of periodically organised, highly similar, structural motifs it is described as a three dimensional lattice where each lattice point is a motif.

In mathematical description, because the X-ray is characterised as a wave, which can be described by a periodic function, a simple wave in one dimension is written in the form

$$\boldsymbol{f(x)} = F_h \cos 2\pi(hx + \alpha) \quad \text{(eq. 2)}$$

Where $f(x)$ specifies the vertical height of the wave at any horizontal position x, F_h the amplitude of the wave, h its frequency, and α its phase.

This equation can also be written in complex form

$$\boldsymbol{f(x)} = F_h[\cos 2\pi(hx) + \ i \ \sin 2\pi(hx)] \qquad \text{(eq. 3)}$$

or

$$\boldsymbol{f(x)} = \ F_h e^{2\pi i(hx)} \qquad \text{(eq. 4)}$$

with $i = (-1)^{1/2}$.

To describe a complicated wave in one dimension, the sum of simple waves is used. This sum is called a Fourier sum and each simple wave equation in the sum is called a Fourier term

$$\boldsymbol{f(x)} = \sum_{h=0}^{n} F_h \cos 2\pi(hx + \ \alpha) \qquad \text{(eq. 5)}$$

or

$$\boldsymbol{f(x)} = \sum_{h=0}^{n} F_h \, e^{2\pi i(hx)} \qquad \text{(eq. 6)}$$

Although the equation does not show the phase α of the wave, this value is implicit in the combination of the cosine and sine functions and specified by the values of h and x.

The Fourier sum is also applied to a three dimensional wave which has its own three frequencies h, k, and l in each dimension x, y and z respectively

$$\boldsymbol{f(x,y,z)} = \sum_{h}\sum_{k}\sum_{l} F_{hkl} e^{2\pi i(hx+ky+lz)} \qquad \text{(eq. 7)}$$

Like the wave in one dimension, for each possible set of values h, k, and l, the associated wave has amplitude F_{hkl} and the implicit phase α_{hkl}.

In crystallography, each atom j at a position (x_j, y_j, z_j) in a unit cell will give a scattering ray, called an atomic structure factor f_j, which has its own amplitude f_{hkl} and frequencies h, k, and l in the three directions x, y and z or also the indices of specific reflection in the reciprocal lattice

$$\boldsymbol{f_j} = f_{hkl} e^{2\pi i(hx+ky+lz)} \qquad \text{(eq. 8)}$$

Since the unit cell represents all the information of a crystal, diffraction data of a crystal received at the detector is a collection of all diffracted X-rays from all atoms in the unit cells. Each diffracted ray is a complicated wave, the sum of diffractive contributions from all atoms in the unit cell. For a unit cell containing n atoms, the structure factor F_{hkl}, which describes the diffracted ray producing a reflection with reciprocal lattice indices hkl on the detector, is the sum of all the atomic scattering $\boldsymbol{f_j}$ values in a unit cell in the direction defined by h, k, and l

$$\boldsymbol{F_{hkl}} = \sum_{j=1}^{n} f_{hkl}\, e^{2\pi i(hx+ky+lz)} \qquad \text{(eq. 9)}$$

In general, the aim of crystallography is to obtain the mathematical function whose graph is the molecular electron density map, enabling interpretation of the desired structure. X-rays are scattered from the electrons of all the atoms of the crystal. The diffraction obtained is thus a representation of the clouds of electrons in the molecules of the crystal, called electron density (Rhodes, 2006). In other words, $\boldsymbol{F_{hkl}}$ can be written as the sum of contributions from each volume element of electron density in the unit cell. By making the size of the volume element infinitesimally small, we obtain in the limit the following integral

$$\boldsymbol{F_{hkl}} = \int_x \int_y \int_z \rho(x,y,z) e^{2\pi i(hx+ky+lz)}\, dx\, dy\, dz \qquad \text{(eq. 10)}$$

or

$$\boldsymbol{F_{hkl}} = \int_V \rho(x,y,z) e^{2\pi i(hx+ky+lz)} dV \qquad \text{(eq. 11)}$$

where the integral over V, the unit cell volume, is just shorthand for the integral over all values of x, y, and z in the unit cell. Each volume element contributes to F_{hkl} with the phase determined by its coordinates (x, y, z), just as the phase of atomic contributions depend on atomic coordinates. This equation means that F_{hkl} is the Fourier transform of $\rho(x, y, z)$ on the set of real lattice planes (h, k, l) (Rhodes, 2006). Since the Fourier transform is reversible, equation (eq. 11) can be written as

$$\rho(x,y,z) = \frac{1}{V} \sum_h \sum_k \sum_l \boldsymbol{F_{hkl}} e^{-2\pi i(hx+ky+lz)} \qquad \text{(eq. 12)}$$

This equation is a sum rather than an integral because $\boldsymbol{F_{hkl}}$ represents a set of discrete functions.

Considering $\boldsymbol{F_{hkl}}$ as a vector, $\boldsymbol{F_{hkl}}$ can be written in the form

$$\boldsymbol{F_{hkl}} = |\boldsymbol{F_{hkl}}| e^{i\alpha_{hkl}} \qquad \text{(eq. 13)}$$

where α_{hkl} is the phase angle of the complex structure factor $\boldsymbol{F_{hkl}}$.

Equation (eq. 12) can thus be rewritten as

$$\begin{aligned} \rho(x,y,z) &= \frac{1}{V} \sum_h \sum_k \sum_l |\boldsymbol{F_{hkl}}| e^{i\alpha_{hkl}} e^{-2\pi i(hx+ky+lz)} \\ &= \frac{1}{V} \sum_h \sum_k \sum_l |\boldsymbol{F_{hkl}}| e^{-2\pi i(hx+ky+lz-\alpha'_{hkl})} \end{aligned} \qquad \text{(eq. 14)}$$

with $\alpha'_{hkl} = 2\pi\alpha_{hkl}$

From equation (eq. 14), the electron density can be obtained from $\boldsymbol{F_{hkl}}$, which is characterised by its own amplitude $|\boldsymbol{F_{hkl}}|$ and phase α_{hkl} at each set of h, k, and l.

However, only reflection amplitudes can be obtained from the measured intensities of reflections ($|\boldsymbol{F_{hkl}}| = \sqrt{I_{hkl}}$) and no direct information about reflection phases is provided by the diffraction experiment. This is called the "phase problem" in crystallography.

Several methods are used in protein crystallography to determine the phases. Typically, they lead to an initial approximate electron-density distribution in the crystal, which can be improved in an iterative fashion, eventually converging at a faithful structural model of the protein.

1.1.4 Data analysis/structure determination

The primary result of an X-ray diffraction experiment is a map of electron density within the crystal. This electron distribution is usually interpreted in (chemical) terms of individual atoms and molecules after the phase problem is solved. The result of model building is a model of which atoms are in agreement with both the electron density map. The atomic model is 'refined' by varying all model parameters to achieve the best agreement between the observed reflection amplitudes (F_{obs}) and those calculated from the model (F_{calc}). The final structure is validated to check its stereochemical quality before being published on the PDB.

1.1.4.1 X-ray data processing

Two separate pieces of information can be found in the reflections of the diffraction images. The first comes from the geometrical arrangement of the reflections, which gives all the information about the crystal lattice and the symmetry of the crystal. During indexing the spots have to be found, identified with integer numbers (h, k, l = Miller Indices) and the crystal geometry has to be determined accurately so that the intensities can be integrated accurately and the space that has to be modelled later is defined accurately. This step will provide crystallographers with information about the space group and unit cell dimensions. A diffraction experiment involves measuring a large number of reflection intensities. Because crystals have symmetry,

some reflections are expected to be equivalent and thus have identical intensity. The average number of measurements per individual, symmetrically unique reflection is called redundancy or multiplicity. Because every reflection is measured with a certain degree of error, the higher the redundancy, the more accurate the final estimation of the averaged reflection intensity. The spread of individual intensities of all symmetry-equivalent reflections, contributing to the same unique reflection, is usually described by the residual R_{merge} (sometimes called R_{sym} or R_{int}) (Wlodawer *et al.*, 2008).

$$R_{\text{merge}} = \frac{\sum_{hkl} \sum_j \left| I_{hkl,j} - \langle I_{hkl} \rangle \right|}{\sum_{hkl} \sum_j I_{hkl,j}} \qquad \text{(eq. 15)}$$

where $I_{hkl,j}$ is the j^{th} intensity measurement of reflection hkl, and $\langle I_{hkl} \rangle$ is the average intensity from multiple observations.

If the value of the overall R_{merge} is too high, the data should be truncated to lower resolution. This value should be smaller than 10 %**.** The intensities of the reflections are the actual experimental data of a crystallography experiment. Therefore integration is a crucial step during data processing. An intensity value (I) and a background value (σI) are determined and saved. Several errors can occur and have to be accounted for: reflections can be hidden in the background, the signal can be saturated and reflections can be hidden by the backstop of the detector. Badly measured reflections should be excluded from the analysis. This is generally done using automatic integration programmes.

The second comes from the intensity of the reflection which gives information about the contents of the crystal. Unfortunately the second part of information, which is the one that we are actually interested in, is only partial - we lack the phases.

The final step of data processing is scaling. During scaling the integrated values of the different images collected during the diffraction experiment are combined into one set of structure factors and normalised, also according to symmetry.

1.1.4.2 Phase determination

1.1.4.2.1 Experimental phasing

The phase problem was conventionally solved using additional experimental information from a number of derivatives of the native crystals which were made by soaking in one or more solutions containing heavy atoms such as Hg, Pt, and Au. The additional scattering of the heavy atoms results in a difference in intensity of the observed reflections, which is exploited to obtain phase estimates. This approach is known as Multiple Isomorphous Replacement (MIR).

A major drawback of this method is that the derivative crystals may have different unit cell dimensions compared to the native crystal. This obstacle is overcome by the Single or Multi-wavelength Anomalous Diffraction (SAD or MAD) method, although the anomalous signals obtained from these methods are much smaller than the isomorphous signal from the MIR method. These methods provide one or more datasets from the same crystal containing suitable anomalous scatterers. The phases are calculated from the wavelength-dependent quantitative differences in the anomalous scattering contribution of certain atoms contained in the crystal. In anomalous scattering methods, the intensity differences between Friedel pair reflections hkl and –h-k-l (the so-called Bijvoet differences) are used to calculate phase estimates. The continuous tunability of synchrotron radiation sources makes them convenient to exploit yet another signal: dispersive intensity differences between data collected at different wavelengths.

1.1.4.2.2 Direct methods

Direct methods have been developed for solving the phase problem in the structure determination of small molecules, and have also used in protein structure determination as an assistive tool complementary to isomorphous replacement, and anomalous diffraction. However, this technique is limited to use with small proteins and high resolution data (up to 1.2 Å).

1.1.4.2.3 Molecular replacement

A much more common way of phasing in protein crystallography is molecular replacement (MR). The MR method has been based on the properties of the

Patterson function which corresponds to a map of position vectors between each pair of atoms in the structure. The Patterson map is a vector map, with peaks at the positions of vectors between atoms in the unit cell. The vectors in the Patterson map can be divided into two categories, including intramolecular vectors (from one atom in the molecule to another atom in the same molecule) and intermolecular vectors (from one atom in the molecule to another atom in the other molecule). The intramolecular vectors depend only on the orientation of the molecule, not on its position in the unit cell, while the intermolecular vectors depend on both the orientation of the molecule and its position. The MR process typically consists of two steps: rotation and translation, using a known homologous structure as a starting model to determine the location of the target protein in the unit cell. The rotation function finds the orientation of the reference molecule in the target unit cell by exploiting the intramolecular vectors. Once the orientation of the reference molecule is known, the translation function then finds its position by exploiting the intermolecular vectors. The phases of the correctly placed model are used as starting phases for map reconstruction. This is a quick and common technique in protein structure determination, but model phase bias can be substantial because the phases dominate the electron density reconstruction (Rupp, 2010). Replacement is to be understood as "positioning" of the search probe in the crystal structure, not as "substitution".

In practice, the program Phaser (McCoy *et al.*, 2007) is a popular tool for MR which requires an mtz file and a reference model. An mtz file contains necessary information, such as magnitude of each reflection, symmetry operations, and cell dimensions. A reference model should share at least 30 % sequence identity with the target structure. The user needs to provide the number of molecules in an asymmetric unit, the molecular weight or protein sequence, and percentage sequence identity with the model. The result includes a log file, a new mtz file and an initial structure of the target protein. Phaser uses a Log Likelihood Gain (LLG), the difference between the likelihood of the model and the likelihood calculated from a Wilson distribution as the scoring function (McCoy *et al.*, 2007). The LLG can be used to compare different models against the same data set, which higher LLG better model. A success of MR is evaluated by a LLG of 40 or greater. In addition, the solution is

also judged using the Z-scores, including the rotation function Z-score (RFZ) and the translation function Z-score (TFZ), which indicate the number of standard deviations above or below the mean for particular LLG score. The Z-score is useful in the case that there are many ambiguous solutions found due to a low signal-to-noise value of the search. An acceptable solution should have TFZ above 5.

1.1.4.3 Model building

After an electron density map has been obtained from initial phasing and density modification techniques, interpretation of this map in terms of a protein model is required. In this process prior knowledge of the amino-acid sequence as well as the known structural characteristics of protein molecules are of great importance. An atomic model of the structure has to be built into the electron density map. The correct building of the structure and its refinement is used to improve the (approximate) phases obtained earlier.

1.1.4.4 Refinement

After the initial phasing and building, the model of a protein is generally far from perfect. To improve the phases and also the interpretation of the electron density map, refinement methods are a very important step in the interpretation of the diffraction data. After refinement additional rebuilding rounds are normally needed.

Refining is achieved through adjustment of the atomic coordinates to fit the diffraction data better. The first quality indicator for the structure is the R-factor which measures how the calculated diffraction by the structure fits to the observed intensity data. For a group of reflections h, the R-factor is described as a ratio

$$R = \frac{\sum_h \big||Fobs| - |Fcalc|\big|}{\sum_h |Fobs|} \qquad \text{(eq. 16)}$$

where $|Fobs|$ are the observed structure factor amplitudes and $|Fcalc|$ are amplitudes calculated from the current model. The R-factor falls towards zero as the observed and calculated structure factor amplitudes agree more closely. The progress of refinement can be analysed by monitoring the R-factor and ensuring that it

continues to fall. However, the refinement procedure may make adjustments which reduce the value of R-factor, by modelling the noise in data, without any improvement of the model. To avoid this problem, a factor called R_{free} is used. R_{free} is calculated using the same equation as the R-factor, but it is calculated from a small fraction of reflections (typically 5%) that are excluded from refinement. The reduction of R_{free} is thus an unbiased estimate of the improvement of the model. The R_{free} is typically about 1.2 times the R-factor.

Two methods are widely used in refinement: Non-crystallographic symmetry restraints (NCS) and simulated annealing. Both methods use restraints to how an atomic model should behave, for example bond distances, angles and torsions and temperature factors (B-factors). In NCS each atom of equivalent molecules in an asymmetric unit is treated using the same restraints, increasing the data-parameter ratio (Kleywegt, 1996, Headd *et al.*, 2014). There are however some local conformational differences among equivalent molecules of the model, such as disordered loops and active site regions; manually checking and building each residue of the protein model still necessary. In simulated annealing the structure is "heated" to allow all atoms of the structure rearrange themselves randomly in a liquid phase and then is cooled gradually. During cooling stage, all atoms arrange themselves again to obtain their lowest energy state. This process can improve both the accuracy and variability of the final refined structure (Brunger and Adams, 2002).

1.1.4.5 Validation

A model stays a model even if at high resolution and if it fits the electron density very well. There are many potential sources of error (experimental or due to wrong interpretation) during the structure solution process. Validation methods detect inconsistencies in the final model based on information that was not used during the refinement process, making validation a compulsory step in protein structure determination.

At the same time validation thus can be seen as an additional step of the refinement process. This additional step can be done during final refinement steps if the

validation indicates that something really does not seem right with the structure, which is a good suggestion to improve the structure quality.

When a structure is submitted to a database, an additional validation process is performed. The Protein Data Bank suggests some common programs to validate the structures before submission, such as MolProbity (Chen *et al.*, 2010), which is useful for checking geometrical criteria such as torsion angle, side chain rotamers, and C_β deviations; Procheck (Laskowski *et al.*, 1993), which works on the stereochemical quality of protein structure; and other programs with similar functions.

1.2 Glycosyltransferase

Glycosyltransferases (GTs, Enzyme Commission number (EC) 2.4.x.y) are enzymes responsible for the synthesis of glycoconjugates by transferring an activated sugar residue from a donor either to an appropriate acceptor molecule or to an aglycone for chain initiation and elongation. The donors can be nucleotide-sugars, lipid phosphate sugars or phosphate sugars and the acceptors can be lipids, proteins, heterocyclic compounds, or other carbohydrate residues. There are three classification systems for GTs based on their properties, namely sequences, structures and the stereochemistry of the transferred sugar moiety in the final product. The wide range of donors and acceptors for glycosyltransferases results in significant diversity; enzymes from the same family may have different donors or acceptors, causing inaccuracy or unreliability in GTs function and structure prediction from sequence information. Much research has been carried out to find out more about the relationship between structure and function of GTs which may give insights into GTs evolution as well as their catalytic activities.

The GT family classification available on the Carbohydrate Active enZymes (CAZy) database is based on sequence (Lombard *et al.*, 2014). This system was first proposed by Campbell *et al.* and then modified by Coutinho *et al.* (Campbell *et al.*, 1997, Coutinho *et al.*, 2003). A sequence-based classification spreads GTs in many families, which reflects the diversity of their acceptors and donors. For some researchers, GTs may be grouped based on their common acceptors, for example the galactosyltransferase family, and the sialyltransferase family (Hennet, 2002, Audry *et al.*, 2011).

In contrast, three dimensional structures of GTs are well-conserved. Only two different folds (GT-A and GT-B) have been identified in solved crystal structures up until 2007. A new fold (GT-C) has been reported recently in a structural study of GTs utilising a lipid-phosphate donor substrate (Henrissat *et al.*, 2008).

Additionally, GTs are also classified by the configuration of the anomeric functional group of the glycosyl donor molecule and of the resulting glycoconjugate. All known glycosyltransferases can be divided into two major types: retaining GTs, which

transfer a sugar residue with the retention of anomeric configuration, and inverting GTs, which transfer a sugar residue with the inversion of anomeric configuration (Lairson *et al.*, 2008, Golovin *et al.*, 2005, Kabsch and Sander, 1983).

1.2.1 Classification of glycosyltransferases

1.2.1.1 Based on sequences

GTs have been classified into families by amino acid sequence similarities (Coutinho *et al.*, 2003, Velankar *et al.*, 2005, Campbell *et al.*, 1997) (available at http://afmb.cnrs-mrs.fr/CAZY). At the time of writing (May, 2014), the database comprises more than 140868 known and putative GT sequences that have been divided into 95 families (denoted as GTx). These numbers will be likely to increase with the discovery of new GT genes.

Significant deviations in the number and function of GTs are observed among families. A few families comprise a large number of sequences from various sources with diverse functions. The best known example is the family GT2 which is recorded in the CAZy database with more than 42997 sequences, and members found in a wide range of species including animals, plants, yeasts, and bacteria. These enzymes cover at least 14 distinct GT functions that have already been characterised, such as cellulose synthesis, chitin synthesis, mannose transfer, glucose transfer, galactose transfer, rhamnose transfer, and other functions. In contrast, other families are monofunctional and contain only a few sequences, for example the GT6 family which only catalyse N-acetylgalactosamine (GalNAc) or galactose (Gal) transfer (Breton *et al.*, 2006).

The sequence-based classification system is supposed to incorporate both structural and mechanistic features within each family (Coutinho *et al.*, 2003), however, the system is only well applied to glycoside hydrolases because they have a good correlation of sequence with enzyme mechanisms (inverting or retaining) and once established for a member of a family, the mechanism can be safely extended to all other members of that family (Davies and Henrissat, 1995). In contrast it is tricky to apply this to the GT families due to similarities at the sequence level between the inverting and retaining families (Franco and Rigden, 2003, Liu and Mushegian,

2003). Different families which have different catalytic mechanisms (retaining or inverting) may have similarities in their sequence, such as the retaining GT27 family (related to the animal polypeptide-α-GalNAc transferase) and the inverting GT2 family (Breton *et al.*, 1998a). In addition, enzymes in one family can be either inverting or retaining. For example, the GT52 family, which comprises both inverting (α2,3-sialyltransferase) and retaining (α2-glucosyltransferase) enzymes.

The prediction of the function of a putative GT based on sequence similarity is also difficult because there are many examples of closely related sequences having different catalytic activity, even within a monofunctional family. A good example are histo blood group A and B transferases (belonging to the GT6 family), which differ by only four amino acids but utilise different glycosyl donors UDP-GalNAc (uridine diphospho N-acetyl galactosamine) and UDP-Gal for the A transferase and the B transferase respectively (Yamamoto *et al.*, 1995). Consequently, it is unreliable to predict the functions of members belonging to a large polyfunctional family based on sequence.

1.2.1.2 Based on structures

According to the CAZy database, GTs are classified into 94 different families based on sequence interrelationships (Coutinho *et al.*, 2003). There are, however, only two general folds (GT-A and GT-B) and a proposed fold (GT-C) found in the published structures of GTs at present. The GT-A fold, first observed in the SpsA from *Bacillus subtilis* (Charnock and Davies, 1999), and the GT-B fold, first reported in β-glucosyltransferase structure (Breton *et al.*, 2006), adopt Rossmann-like folds that are typical of nucleotide-binding proteins.

The general characteristics of a GT-A structure comprises an open twisted β-sheet surrounded by α-helices on both sides which are tightly associated, and hence appear as one domain with a continuous central β-sheet (Figure 7A). They also contain an AspXaaAsp motif (DXD motif) which is linked to the metal-dependent catalytic mechanism of most glycosyltransferases. However, this motif is not a signature of either the GT-A fold or glycosyltransferases because there are some non

glycosyltransferases which also contain this motif and some GT-A fold enzymes do not have this motif in their sequence (Lairson *et al.*, 2008).

On the other hand, the GT-B fold consists of 2 β/α/β domains which interact and face each other with the active-site being located in the resulting cleft, and a bound metal ion, which is not essential for activity (Figure 7B). The new fold, GT-C, was found recently in the structure of oligosaccharyltransferase STT3 from *Pyrococcus furiosus*, which adopts a novel structure with a central, mainly α-helical domain surrounded by three β-sheet-rich domains (Figure 7C) (Lairson *et al.*, 2008, Igura *et al.*, 2008).

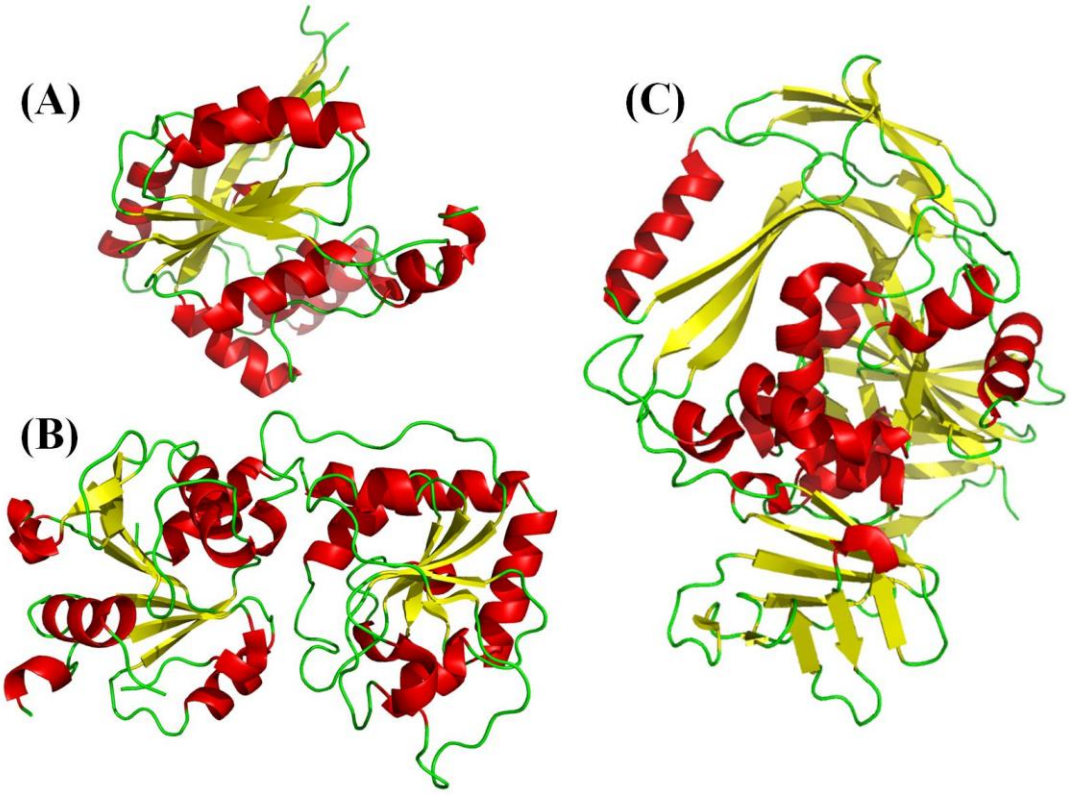


Figure 7. Three different folds of glycosyltransferases. (A) GT-A fold, displayed by SpsA from *Bacillus subtilis* (PDB ID 1QGQ) (Charnock and Davies, 1999), (B) GT-B fold, presented by bacteriophage T4 β-glucosyltransferase (PDB ID 1BGT) (Vrielink *et al.*, 1994, Coutinho *et al.*, 2003), and (C) GT-C fold , exhibited by the C-terminal soluble domain of *Pyrococcus furiosus* STT3 (PDB ID 2ZAG) (Igura *et al.*, 2008). All structures are shown in cartoon representation and are coloured by secondary structure with helices in red, sheets in yellow and loops in green. The picture was created using Pymol.

1.2.1.3 Based on the anomeric configuration of glycosyltransferase product

Based on the configuration of the product, GTs are divided into 2 classes: retaining GTs which retain the anomeric configuration of the monosaccharide from the donor in the product and inverting GTs which invert the stereochemistry of the sugar moiety in the product.

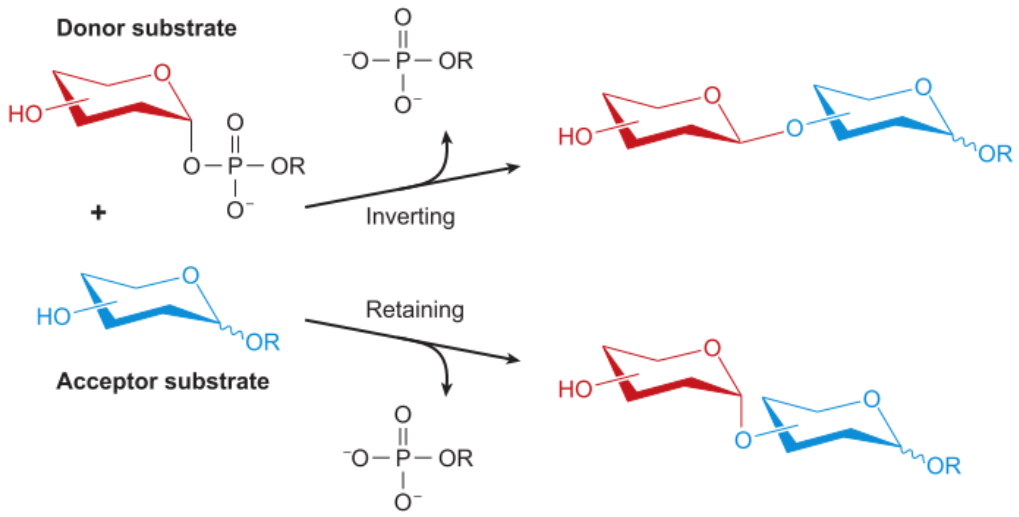


Figure 8. Scheme of either inversion or retention of the anomeric stereochemistry with respect to the donor sugar (Lairson *et al.*, 2008).

1.2.1.3.1 Inverting glycosyltransferases

The GT-A fold inverting GTs cause an inverted stereochemistry of the anomeric moiety of the donor via a single oxocarbenium ion-like transition state by using a single displacement S_N2-like reaction (Lairson *et al.*, 2008). In these enzymes, an active-site side chain serves as a base catalyst to deprotonate the incoming nucleophile of the acceptor, allowing direct S_N2-like displacement of the activated phosphate leaving group (Figure 9). The catalytic base residue is either Asp or Glu, the position of which is conserved in the catalytic site of enzymes from GT2 (SpsA) (Charnock *et al.*, 2001), GT7 (β4-galactosyltransferase (Gal-T1)) (Gastinel *et al.*, 1999), GT13 (N-acetylglucosaminyltransferase (GnT-I)) (Unligil *et al.*, 2000), GT14 (leukocyte type core 2 β1,6-N-acetylglucosaminyl-transferase (C2GnT-L)) (Pak *et al.*, 2006), GT31 (β1,3-*N*-acetylglucosaminyltransferases (Mfng) (Jinek *et al.*, 2006) and GT43 (glucuronyltransferase (GlcAT-P)) (Kakuda *et al.*, 2004, Ohtsubo *et al.*,

2000) families. The importance of these residues was demonstrated by mutagenesis studies of *Sinorhizobium meliloti* glucosyltransferase ExoM from GT2, in which activity of the enzyme was abolished when the Asp187 (positioned at the conserved location in the active site) was mutated (Garinot-Schneider *et al.*, 2000). Inverting GTs in general promote catalysis by features that help to promote leaving-group departure. In the GT-A fold GTs, a DXD motif-bound divalent metal ion is typically positioned to interact with the diphosphate moiety of the sugar nucleotide donor. The additional negative charge that develops on the UDP leaving group during bond breakage is electrostatically stabilised by the positively charged metal ion. The metal ion serves as an acid catalyst initiating a sequential ordered mechanism in which nucleotide sugar binding is followed by loops closing and acceptor binding (Zhang *et al.*, 2001). This motif is observed in a majority of GT-A fold GTs which require divalent metal ions such as Mg^{2+} or Mn^{2+} for their activity. However, there are some exceptions which are metal independent in catalytic mechanism, namely sialyltransferases from GT42 and C2GnT-L from family GT14. These enzymes use tyrosyl hydroxyls or basic amino acids to compensate for the positive charge of the metal ion. This mechanism is also observed in the GT-B fold GTs, explaining their metal-independent activities.

1.2.1.3.2 Retaining glycosyltransferases

Unlike inverting GTs, the catalytic mechanism of retaining GTs is not clearly understood. There are two proposed mechanisms, the double-displacement mechanism and the internal return (S_Ni-like) mechanism (Figure 10). The first mechanism involves formation of a covalent intermediate with inversion of the anomeric configuration of the donor sugar followed by hydrolysis of the intermediate with another inversion (Gastinel *et al.*, 2001). However, neither structural nor kinetic data agreed with this mechanism (Lairson *et al.*, 2004, Boix *et al.*, 2001). Even though many experiments using mutant enzymes are able to identify a range of intermediates, the only intermediate that could be trapped was an amino acid distant from the active site (Lairson *et al.*, 2004). In addition, based on the structural data, only a few retaining GTs have amino acids which have their side chains suitably positioned in the active site to act as a nucleophile in such a mechanism.

An alternative mechanism is an S_Ni-like mechanism in which the nucleophile attacks from the same face as the leaving group departs. This mechanism was first proposed for retaining glycosyltransferases by Persson *et al.* (2001) when they studied α-Galactosyltransferase LgtC from *Neisseria meningitidis* in which stable donor and acceptor substrate analogues were observed bound to the active site. The researchers suggested that a S_Ni-like mechanism occurred involving a direct attack on the C1 atom of UDP-Gal by the acceptor molecule itself concurrent with cleavage of the glycosidic bond and direct transfer of the Gal moiety to the disaccharide acceptor without the formation of a glycosyl enzyme intermediate. The enzyme, therefore, was proposed to orient substrates in close proximity, stabilise the oxocarbenium ion-like species and activate the leaving group, leading to the reduction of transition state energy. The theory was supported recently through the structural and kinetic study of trehalose-6-phosphate synthase (OtsA). In this study, the ternary complex of OstA and UDP-Glc suggests the transition state occurs when a hydrogen bond is formed between the leaving group oxygen of UDP and the nucleophile mimic of the sugar moiety. In addition, the results of kinetic isotope effect experiments and the linear free energy relationships of a range of substrates implicate both the leaving group of the donor and the acceptor nucleophile during the transition state and suggest a front-side, S_Ni-type mechanism (Lee *et al.*, 2011).

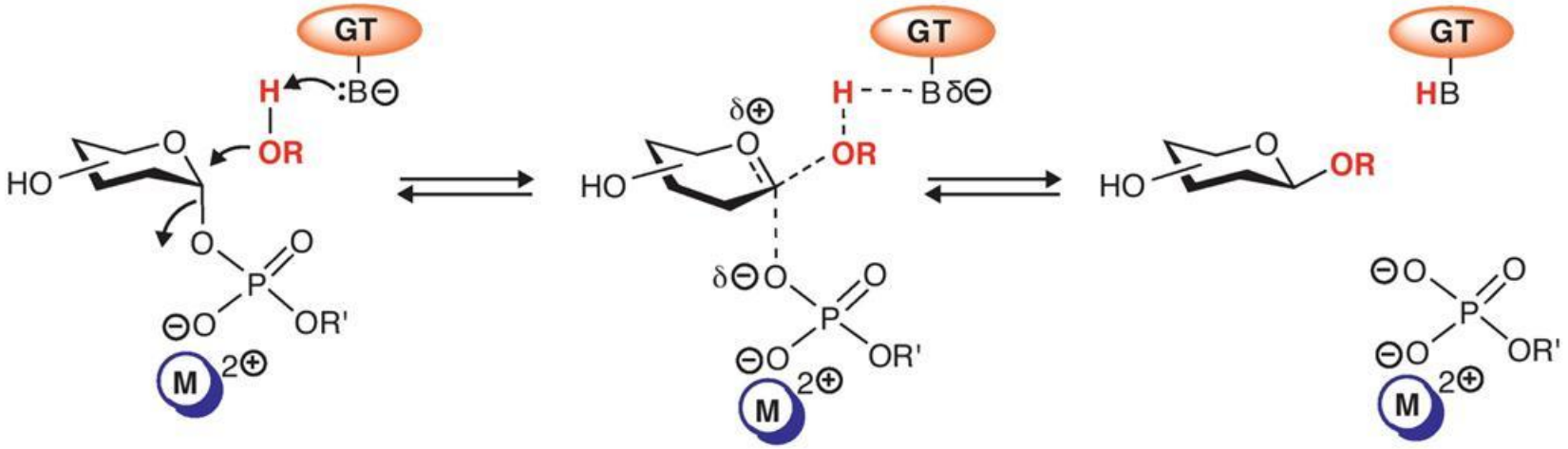


Figure 9. Reaction mechanism proposed for inverting glycosyltransferases. Inverting GT reactions are suggested to occur in single displacement with formation of an oxocarbenium-ion transition state. A catalytic amino acid serves as general base (noted as B) that deprotonates the nucleophile OH-group of the acceptor (HOR) and the negative charge on the departing phosphate can be stabilized by metal (noted as M) as shown in the figure (for metal binding GT-A enzymes) or positive amino acids or helix dipole (for GT-B enzymes) (Breton *et al.*, 2012).

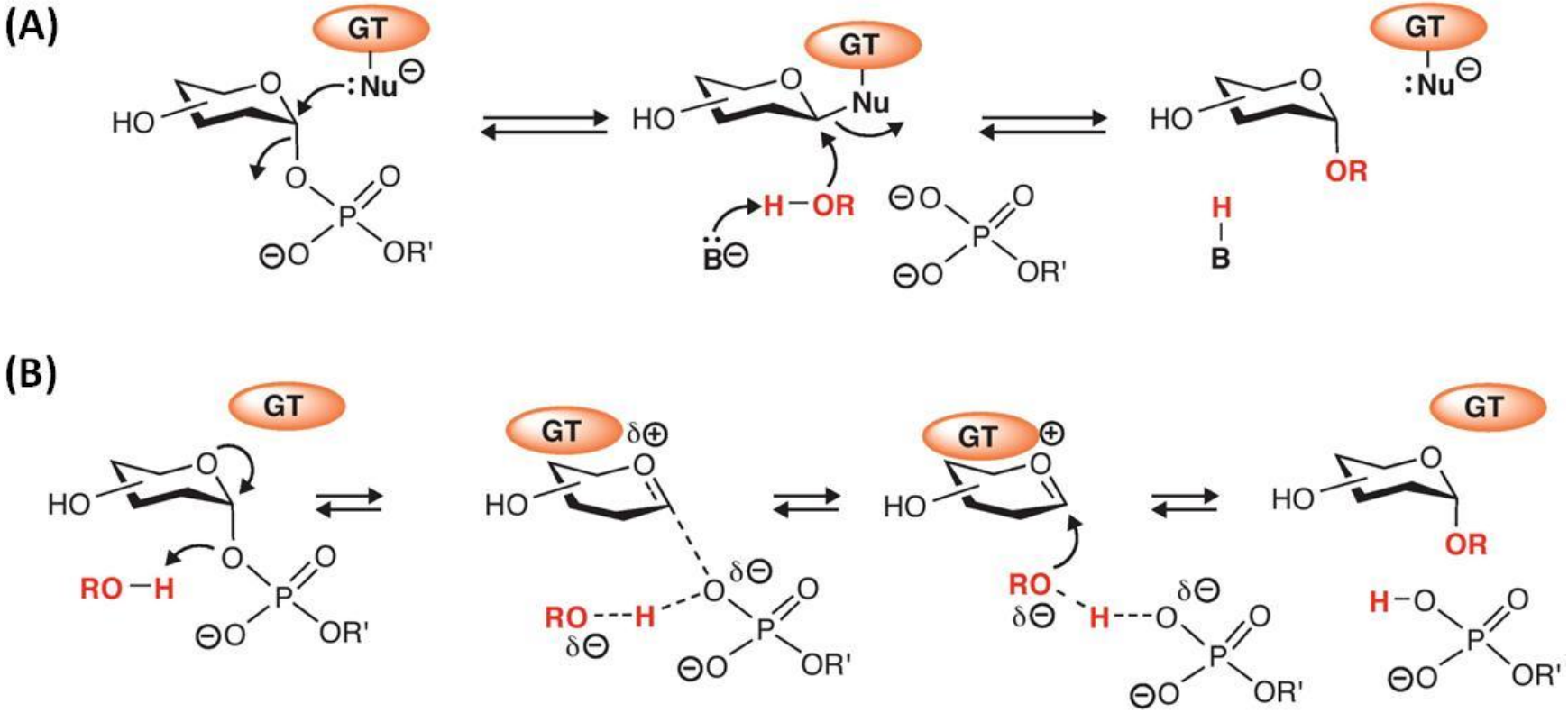


Figure 10. Reaction mechanism proposed for retaining glycosyltransferases. (A) illustrates the double displacement mechanism. (B) illustrates the internal return (S_Ni-like) mechanism. ROH is noted for an acceptor and R′ an active donor (Breton *et al.*, 2012).

1.2.2 Glycosyltransferase family 6

Glycosyltransferase family 6 consists of α-1,3-galactosyltransferase (EC 2.4.1.87); α-1,3 N-acetylgalactosaminyltransferase (EC 2.4.1.40); α-galactosyltransferase (EC 2.4.1.37) and globoside α-N-acetylgalactosaminyltransferase (EC 2.4.1.88) (CAZy database). These enzymes catalyse the stereochemistry-retaining transfer of α-galactose (αGAL) or N-acetylgalactosamine (GalNAc) from a specific donor UDP-α-D-galactose (UDP-Gal) or UDP-*N*-acetyl-D-galactosamine (UDP-GalNAc) to the 3-OH group of a β-linked Gal or GalNAc of an acceptor. This family was researched through their representatives, including the histo-blood group A and B GTs (GTA and GTB), the α1-3-galactosyltransferase (α3GT), Forssman α1-3 GalNAc-transferase (FS), isogloboside 3 synthase (iGb3S) and their homologues from other vertebrates and prokaryotes (Hennet, 2002).

1.2.2.1 Functions

This family is interesting due to the antigenic effects of their products on human immune systems. The most well characterised GT6 members are GTA and GTB, which are responsible for the ABO blood group system classification. In humans, the ABO locus located on chromosome 9 consists of 3 alleles, A, B and O. GTA and GTB are encoded by A and B alleles at the ABO locus respectively. Their respective products, A or B histo-blood group antigens, are formed by the transfer of a Gal or GalNAc to the Gal moiety of their specific acceptor H antigen, which is recognised with a fucose molecule attached to the C2 position of the Gal moiety (Figure 11). Non-functional products of various O alleles are not able to catalyse Gal/GalNAc transfer to H antigen to produce either A or B histo-blood group antigen (Yamamoto, 2000, Yamamoto *et al.*, 1990). In nature, there exist at least four H antigens on glycolipids and glycoproteins that are recognised by GTA and GTB. The most common are the type I H antigen (α-L-Fucp-(1→2)-β-D-Galp-(1→3)-β-D-GlcNAcp-OR where R as either glycoprotein or glycolipid) and the type II H antigen (α-L-Fucp-(1→2)-β-D-Galp-(1→4)-β-D-GlcNAcp-OR). Less common are the type III H antigen (α-L-Fucp-(1→2)-β-D-Galp-(1→3)-α-D-GalNAcp-OR) and the type IV H antigen (α-L-Fucp-(1→2)-β-D-Galp-(1→3)-β-D-GalNAcp-OR) (Hakomori, 1999) (Figure 11).

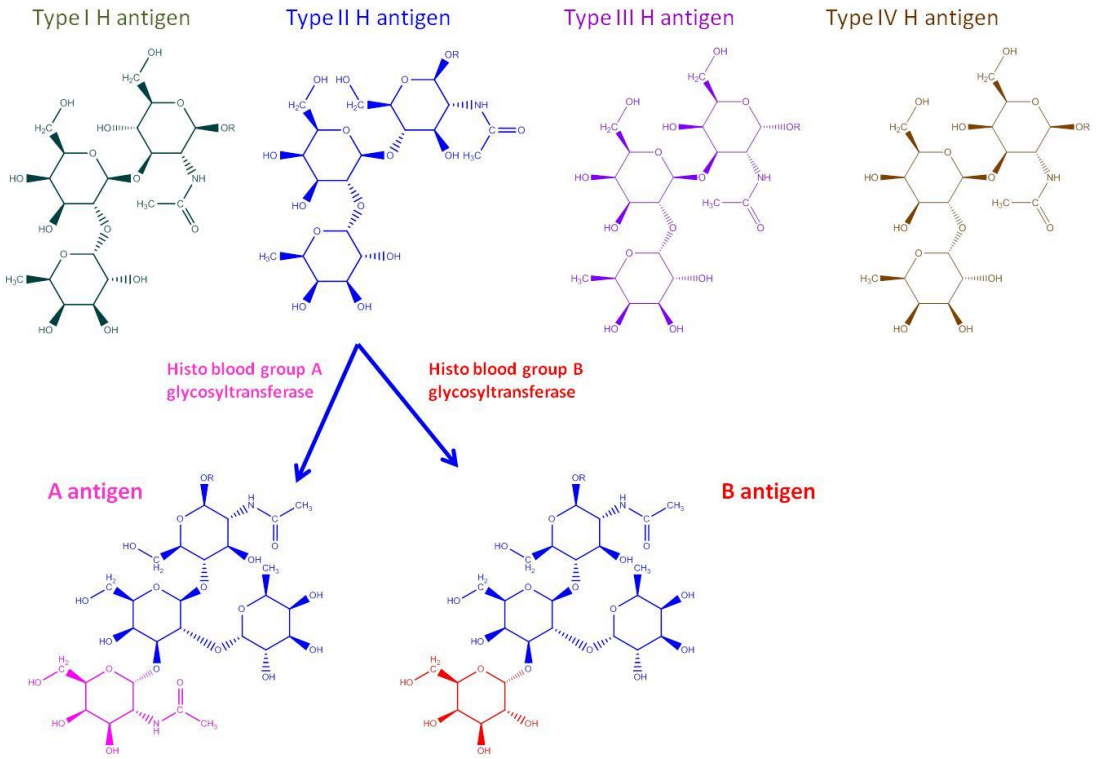


Figure 11. Chemical structure of 4 types of H antigen acceptors of the human ABO(H) Blood Group A and B Glycosyltransferases. Picture created using ChemDraw.

Understanding the ABO polymorphism in humans prevents many fatal transfusions from occurring due to incompatible blood transfusion. ABO variety also helps human immune system resist bacterial and viral pathogens. ABO glycans were detected on membrane protein 120 of HIV derived from cells transfected with ABO cDNA. Virus carrying the A histo-blood group antigen was neutralised in fresh human sera from B and O histo-blood group cells and *vice versa* (Neil *et al.*, 2005). This finding suggests that interaction between ABO sugars and the envelope of HIV is a barrier to HIV transmission between individuals of different ABO types. A similar result was also recorded with the measles virus (Preece *et al.*, 2002). Antibodies of ABO antigens can inhibit the transmission of Severe Acute Respiratory Syndrome coronavirus (SARS-coV) by interrupting the interaction between the glycosylated SARS-coV spike (S) protein and angiotensin-coverting enzyme 2 as its receptor on the host cell surface (Guillon *et al.*, 2008). Both the presence of ABO antigens on virus envelopes and the protection of ABO antibodies

against virus transmission were reported as two main driving forces in ABO genetic evolution (Seymour *et al.*, 2004).

Unlike ABO genes of which both active and inactive products are produced in humans, the α3GT, FS and iGb3S genes do not encode any protein because of frameshift and missense mutations (Koike *et al.*, 2002).

α3GT is an enzyme that catalyses the transfer of galactose from UDP-α-D-Galactose as donor substrate to the acceptor N-acetylglucosamine forming an α-1,3 linkage with β-galactosyl groups in glycoconjugates, also known as α-Gal epitopes. The enzyme is neither expressed in humans nor in apes and old world monkeys (Galili *et al.*, 1987, Galili *et al.*, 1988). Only partial sequences, similar to the α3GT gene of non-primate mammals have been detected in the human genome and are thought to correspond to pseudo genes, due to multiple deletions leading to premature translation stops (Lanteri *et al.*, 2002). Complete loss of expression of this gene in old world primates enables the production of significant amounts of natural antibodies toward α-Gal epitopes (Galili *et al.*, 1987). These antibodies are strongly active in the presence of complement, which is thought to be an effective way to prevent infection by pathogenic microorganisms that expose α-Gal epitopes. However, α-Gal epitopes are also found in many other species, such as pigs, rats, and chickens, and this has been a major barrier against xenotransplantation (Galili, 2001).

FS transfers a GalNAc to a glycolipid, the globotetraosylceramide (Haslam and Baenziger, 1996). From species to species, the Forssman gene seems to be active or inactive. Mouse and chicken are Forssman positive, whereas rat and pigeon are Forssman negative. In nonprimates, Forssman antigen is synthesised by FS. In humans, Forssman glycolipid is not detected, but the precursor globoside is, suggesting that human tissues lack Forssman synthase activity. Forssman glycolipids, normally found only on RBCs of selected nonprimate mammals, are strongly expressed on human A_{pae} RBCs. A_{pae} RBCs, first described in 1987, expresses only part of the normal A antigen and the anti-A present in the serum does not have an anti- A_{pae} component (Stamps *et al.*, 1987). Recently, analysing genetic polymorphisms in the human FS gene (GBGT1) showed that this gene can alter the

enzymatically inactive human protein to its active nonprimate counterpart, implicating the Forssman antigen as a new histo-blood group system with potential implications for contribution to variable host susceptibility to microbial pathogens. (Svensson *et al.*, 2013).

The last functional member of the GT6 family, iGb3S, transfers a Gal on lactosylceramide that is involved in the isoglobo-series glycolipid pathway. This gene is a pseudogene in human which is not able to produce a functional enzyme, but is active in several other mammals (Keusch *et al.*, 2000a). This supports the idea of a selection pressure exerted against the expression of Gal(α-1,3)Gal and GalNAc(α-1,3)Gal antigens in humans. (Iso)globo-series glycolipids often function as receptors for pathogens like bacteria, viruses and toxins (Karlsson, 1995). It is possible that the suppression of Gal(1-3)Gal-related epitopes may give an advantage for the respective hosts toward various microbes and toxins (Hennet, 2002). The rat gene encoding the iGb3 synthase enzyme has been isolated by an expression cloning strategy (Keusch *et al.*, 2000b). The rat iGb3 synthase protein shares about 39% identity with the other α3GalTs and with the Forssman α-1,3-GalNAc transferase (Haslam and Baenziger, 1996).

1.2.2.2 Structures

All characterised vertebrate GT6 members are type II transmembrane proteins localised in the Golgi apparatus, with a cytoplasmic domain, a transmembrane domain of about 20 amino acids, and a C-terminal catalytic domain (Hennet, 2002). The N-terminal cytoplasmic domain of mammalian GTs was reported as a signal for localisation of anchored Golgi GTs (Tu and Banfield, 2010). Although the sequences of GT6 members are diverse, their three dimensional structures are highly conserved and belong to the GT-A fold with an α/β/α Rossmann-like motif which is often observed for nucleotide binding proteins.

The first crystal structure of this family was the catalytic domain (residues 80 – 368) of apo-form bovine α3GT (PDB ID 1FG5) first reported in the tetragonal crystal form (form I), $P4_12_12$, at 2.8 Å by Gastinel *et al.* (Gastinel *et al.*, 2001). The overall structure of the enzyme consists of 10 β-strands, 6 α-helices and 6 3_{10}-helices which

form the GT-A fold with a central core of twisted β-sheet of 8 β-strands surrounded by 4 long α-helices. The central β-sheet can be divided into two main regions which create a cleft between them. The first region runs from Val129 to Met224, defining the N-terminal subdomain and is made up of a β-strand core (3 β-strands) and surrounded by two long α-helices. This subdomain accommodates the nucleotide moiety of the donor substrate binding site. The second portion of the central β-strands consists of 2 parallel β-strands flanked by 2 anti-parallel β-strands with 2 long helices on each side. There is a small β-sheet region parallel to the central β-strand comprising 2 short anti-parallel β-strands. The structures of α3GT in complex with UMP (PDB ID 1G8O) and Hg-UDP-Galactose (PDB ID 1G93) were also reported at 2.3 Å and 2.5 Å respectively (Gastinel *et al.*, 2001).

The higher resolution (1.53 Å) structure of bovine α3GT in complex with UDP and Mn^{2+} was solved in a monoclinic form (form II), $P2_1$, giving a detailed picture of bovine α3GT interacting with its donor substrate (Boix *et al.*, 2001). This structure is similar to that of the form I, which displays Rossmann-like domain in overall structure, but the catalytic site and C-terminus are strikingly different. In general, the structure in form II is more ordered with a 3 times lower average B factor and displays the C-terminal region from residues 357-368 which was disordered in the form I structure. This region acts as a lid over the active site of the enzyme, which is a deep tunnel inside the molecule, with the presence of a UDP moiety and a Mn^{2+} ion (Boix *et al.*, 2001). The UDP moiety orientation is similar to those of UDP-Gal or UMP in form I and also interacts with the Asp225XaaAsp227 motif. These regions underwent a significant conformational change caused by the interactions between the enzyme and the ligand UDP.

The crystal structures of the catalytic domains (residues 63–354) of the cloned GTA and GTB were solved to 1.80 and 1.65 Å resolution, respectively, and of the catalytic domain of the GTA and GTB in complex with the H-antigen disaccharide and UDP to 1.35 and 1.32 Å resolution, respectively (Patenaude *et al.*, 2002). The topology of GTA and GTB resembles the α3GT transferase structure in which the polypeptide chain is organised in two domains separated by a 13 Å cleft containing the active site and all four critical amino acid residues. These four residues are the only difference in sequence between GTA and GTB, namely Arg/Gly 176, Gly/Ser 235, Leu/Met

266 and Gly/Ala 268 respectively. The conserved DXD motif of this family, defined by residues Asp211, Val212 and Asp213, is also found in the active sites of these enzymes (Yamamoto *et al.*, 1990, Patenaude *et al.*, 2002).

Although Forssman antigens were discovered in 1980s and have been intensely researched, the Forssman synthase and iGb3 synthase structures have not yet been reported. Sequence alignment of the 4 GT6 representative enzymes, including Forssman glycolipid synthase (Haslam and Baenziger, 1996), isogloboside 3 synthase (Keusch *et al.*, 2000b), GTA and GTB (Yamamoto *et al.*, 1990), and bovine α3GT (Boix *et al.*, 2001), showed 44% to 55% identity with a few insertions or deletions in the regions corresponding to secondary structures, which suggested they should have similar overall structures to those reported for bovine α3GT, GTA and GTB (Heissigerova *et al.*, 2003). They do however utilise different acceptors; hence, some key residues may play different roles in acceptor binding interactions.

1.2.2.3 Interactions with substrates

Crystal structures of bovine α3GT in substrate free form and substrate bound form show a contiguous flexible loop at the end of the C-terminal region, residues 358 – 368, and an internal loop from residues 188 to 199 that undergoes a conformational change to a more rigid structure when the enzyme interacts with substrates (Boix *et al.*, 2001, Jamaluddin *et al.*, 2007). Like bovine α3GT, the substrate bound complex structures of GTA and GTB also undergo a conformational change related to enzyme-ligand interactions (Qasba *et al.*, 2005, Soya *et al.*, 2009). Previous kinetic and structural studies on the blood group GTs suggest that the UDP-sugar donor substrate binds first to an "open" form of the enzyme (Alfaro *et al.*, 2008, Kamath *et al.*, 1999). Reorganisation of an internal flexible loop (residues 176–188 in GTA/GTB which corresponds to residues 188 – 199 in bovine α3GT) concomitant with donor binding generates a "semiclosed" state and creates an acceptor binding site. Upon binding of the acceptor substrate, the enzyme adopts a "closed" conformation in which the final nine C-terminal amino acid residues (residues 345 – 354 in GTA/GTB corresponding to residues 358 – 368 in bovine α3GT) become ordered by forming hydrogen bonds to both UDP and the acceptor (Alfaro *et al.*, 2008, Boix *et al.*, 2001). Such extensive conformational rearrangements during the

enzyme catalytic cycle are characteristic of GTs in general and have been observed for many other enzymes in this family (Qasba *et al.*, 2005).

Looking into the details of the interactions between residues and substrates, regardless of the diverse donor and acceptor substrate specificity of the members of the GT6 family, the reported structures of the GT6 members display a conserved structure-function relationship. There are nine different amino acid regions containing conserved residues which are categorised based on their roles in substrate interactions (Figure 12) (Heissigerova *et al.*, 2003). LBR-A (residues 133-139, numbered relative to bovine α3GT) interacts with uridine and the ribose ring; LBR-B (residues 199-201) interacts with the ribose ring and Gal or GalNAc of UDP-Gal; LBR-C (residues 225-227, with the conserved DXD motif) interacts with Mn^{2+}, phosphates, and Gal or GalNAc; LBR-D (residues 247-250) interacts with the acceptor substrate; LBR-E (residues 278-282) and LBR-F (residues 314-317) interact with Gal or GalNAc of the donor substrate, LBR-G (residues 340-343), LBR-H (residues 356-361) and LBR-I (residues 365-367) interact with the acceptor and the phosphate atoms of the donor (Figure 13).

LBR-A, LBR-B and LBR-C showed high conservation, especially the DXD motif of LBR-C, which is obviously important in the vertebrate GT6 catalytic activity. The important role of this motif in the GT6 family was shown clearly in structures of the family members in complex with their donor substrate, in which both Asp residues of the DXD motif not only interacted with the Mn^{2+} through coordinate bonds but also the phosphate group (Boix *et al.*, 2001, Patenaude *et al.*, 2002). GTA, GTB and bovine α3GT were all found to bind weakly or not at all to their donor and acceptor substrates in the absence of Mn^{2+} (Soya *et al.*, 2009, Zhang *et al.*, 2001). Mutation of either Asp residue in the DXD motif caused bovine α3GT to be inactive (Zhang *et al.*, 2001).

The flexible LBR-H and LBR-I loops in the C-terminal region contain two conserved residues, Lys359 and Arg365 (numbering relative to bovine α3GT) which were considered to form the lid of the nucleotide sugar binding site (Boix *et al.*, 2001, Heissigerova *et al.*, 2003). As these regions react with the donor substrate on binding, they may play an important role in enzyme activity. In the structures of

bovine α3GT wild type in complex with UDP, both Lys359 and Arg365 make contacts with UDP through phosphate groups (Boix *et al.*, 2001) and their substitutions result in a significant reduction in the enzyme catalytic activity. The activity of the mutants Lys359Arg or Arg365Lys was reduced 30-fold whereas the mutant Lys359Ala or 4 C-terminal end residue truncated form (including Arg365) caused a major loss in enzyme activity (about 350-fold or 150-fold reduction in k_{cat} respectively) (Jamaluddin *et al.*, 2007). However, the affinity of these mutants for the substrate UDP-gal was increased (in case of Lys359Arg and Lys359Ala) or slightly increased (about 2-fold in case of Arg365Lys). This finding suggests that Lys359 and Arg365 are important, although not essential, for activity and appear to have a principal role in transition state stabilisation, mediated through interactions with the UDP leaving group, as opposed to (ground-state) substrate binding.

Two more important regions are LBR-D and LBR-E which contain the only 4 different residues between GTA and GTB. These residues are supposed to be responsible for specificity of the donor recognition where two of the four critical amino acids serve to differentiate between the two sugar-nucleotide donors. In detail, GTB can easily prohibit bulkier A donor (UDP-GalNAc) as the floor of the active site cleft contains the larger critical residues Met 266 and Ala 268 that restrict the space available to UDP-donor sugar moieties. In contrast, the smaller critical residues Leu 266 and Gly 268 in GTA form a larger active site cleft that would be inappropriate for the smaller donor UDP-Gal (Patenaude *et al.*, 2002). The other residues are not positioned to directly interact with the donor. The residue Gly/Ser235 affected the conformation of the aliphatic tail of the acceptor in the structures of GTA and GTB in complex with their UDP-sugar donors and H antigen, suggesting it may play a role in H antigen structure recognition (Patenaude *et al.*, 2002). Interestingly, mutant Arg176Gly of GTA did not show any GTB activity but increased GTA activity although the affinity for the substrates was reduced. Mutant Gly176Arg of GTB showed higher affinity for UDP-Gal than the wild type GTB but the catalytic efficiency was unchanged (Lee *et al.*, 2005, Alfaro *et al.*, 2008).

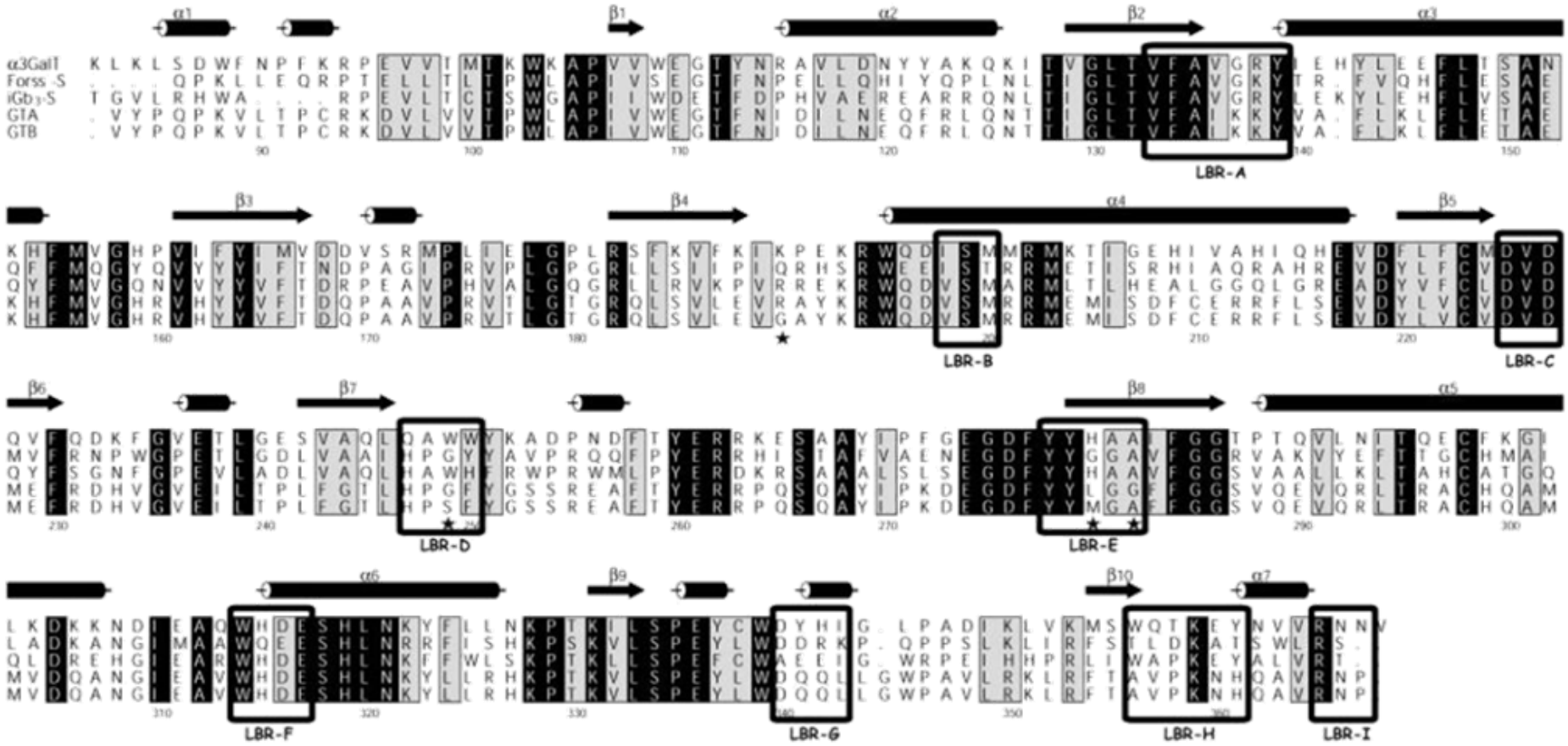


Figure 12. Multiple alignment of amino acid sequences of GT6 family representatives. α3GalT is the bovine α3GalT (GenBank accession number J04989), Forss-S is the canine Forss-S (U66140), GTA is human GTA (J05175), GTB is human GTB (AF134414), and iGB_3S is rat iGb3 (AF248543). Secondary structure elements of α3GalT are indicated above the sequences and numbered as in Boix *et al.* (2001). Identical amino acids have a black background and conservatively-substituted ones a grey background. Regions involved in ligand binding are boxed. The stars indicate the four amino acids that are different between GTA and GTB sequences (Heissigerova *et al.*, 2003).

Structural studies of these mutants showed that Arg/Gly176 had no interaction with either substrate or the other residues of the enzyme and was partially or fully disordered in all structures. This finding, combined with the kinetic assays, suggests that this residue only contributes to substrate turnover in the way smaller side chains with a higher range of dihedral angles can enhance the rate of product release and substrate exchange (Alfaro *et al.*, 2008). In bovine α3GT, mutagenic analysis of His280, corresponding to Met/Leu266 in GTA/GTB, also exhibited the key role of this residue in donor substrate recognition through the H-bond between the residue and the 2-OH of the sugar moiety of the donor substrate (Zhang *et al.*, 2003).

The LBR-F defined as W314H/QD/EE317 (numbering relative to bovine α3GT) is well conserved in the GT6 family. The residues in this region directly contact the donor and acceptor substrates, suggesting that it can play a crucial role in the catalytic mechanism of the GT6 enzymes. Glu317 of bovine α3GT or Glu 303 of GTA/GTB was considered as a catalytic nucleophile because of its close proximity to the C1 of the sugar moiety of the UDP-sugar donor (Gastinel *et al.*, 2001, Patenaude *et al.*, 2002). However, the higher resolution structure of bovine α3GT in complex with UDP displayed an H-bond between Glu317 and 4-OH of the Gal moiety of the acceptor (Boix *et al.*, 2001). Regardless of the debated catalytic mechanism of the GT6 members, the interactions of this residue with the substrate as well as the significant reduction in activity of its mutations such as Glu317Gln, Glu317Ile and Glu317Ala of bovine α3GT or Glu303Ala of GTA/GTB show that the residue is crucial for the enzyme activity (Zhang *et al.*, 2003, Monegal and Planas, 2006, Patenaude *et al.*, 2002).

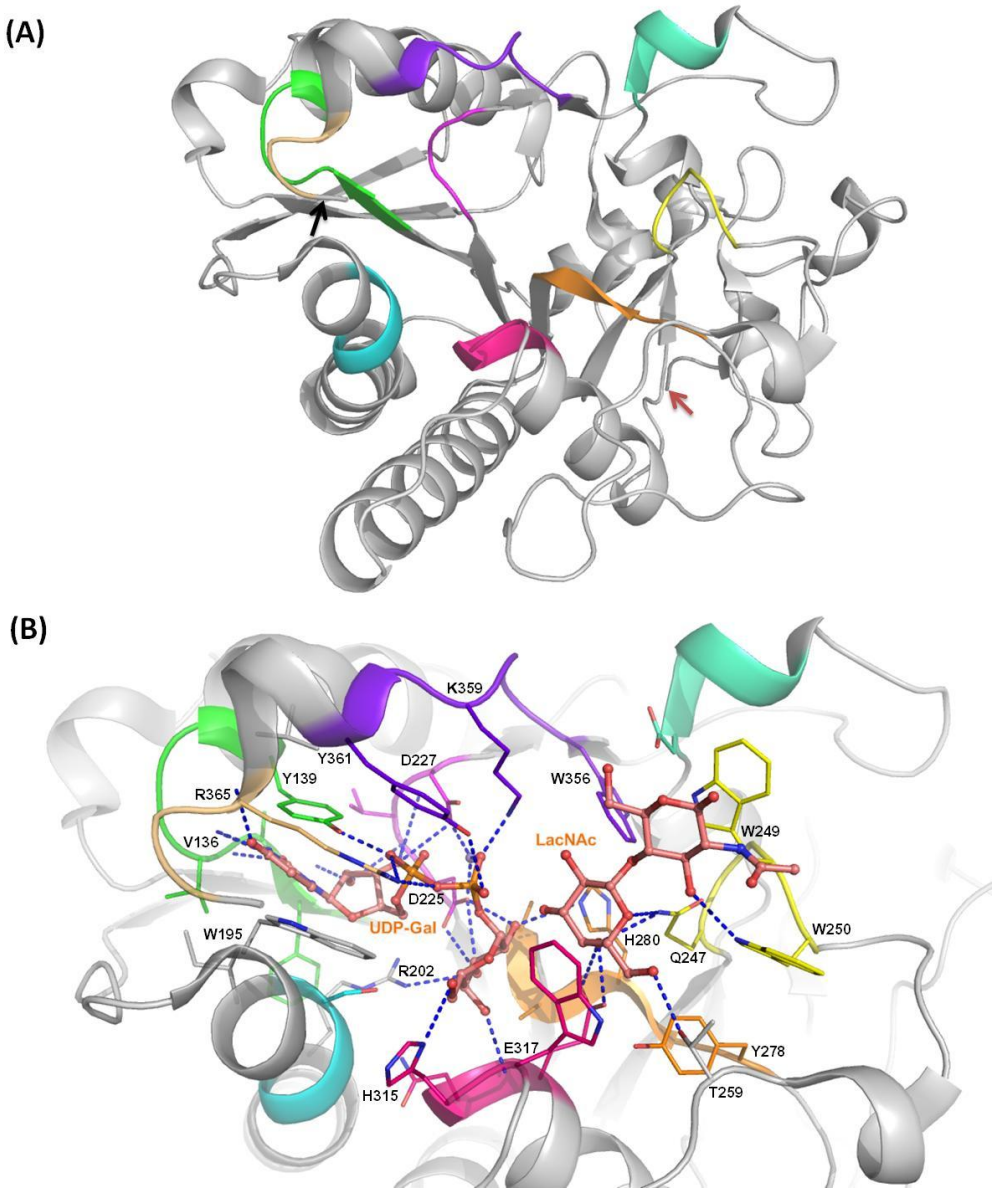


Figure 13. Nine conserved ligand binding regions (LBRs) of GT6a family. (A) Location of nine LBRs on a representative of GT6a family, bovine α3GT (PDB ID 1K4V). LBR-A (residues 133-139) is coloured in green, LBR-B (residues 199-201) in cyan, LBR-C (residues 225-227) in magenta, LBR-D (residues 247-250) in yellow, LBR-E (residues 278-282) in orange, LBR-F (residues 314-317) in hot pink, LBR-G (residues 340-343) in green cyan, LBR-H (residues 356-361) in purple blue, and LBR-I (residues 365-367) brown. Protein is shown in cartoon representation. The black arrow points out the C-terminus and the red arrow the N-terminus. (B) The active site of model bovine α3GT in complex with UDP-Gal and LacNAc. Interacting residues are shown as line, labelled and coloured according to which LBR they belong to. The protein is shown in cartoon representation in grey. The ligands are shown as stick in pink. The figure was created using Pymol.

Kinetic assays, in which the mutant Asp316Asn was inactive while the Asp316Glu mutation reduced the catalytic rate 3-fold, showed that the negative charge of the residue Asp316 (in the case of bovine α3GT) is necessary for catalytic activity because with different side chain size but the same charge, Asp316Glu still retains enzyme activity, while the loss of the hydroxyl group in the case of Asp316Asn resulted in an inactive enzyme (Tumbale *et al.*, 2008). In contrast, both the change in charge (His319Glu) and size (His319Tyr) of His319 ruined the enzyme activity. However, the size of this residue is more important than its charge because the mutant His319Ala activity was only 2-fold reduced compared to the wild type. The size effect was also observed on His315 where the mutant His315Arg resulted in more than 450-fold activity reduction. The impact of these residues on the enzyme activity could reflect their proximity to Glu317 (Tumbale *et al.*, 2008).

1.2.2.4 Catalytic mechanism

Belonging to the retaining GT-A glycosyltransferases, for which the catalytic mechanism is as yet unclear, the enzymatic activities of GT6 members have been greatly investigated. The two postulated mechanisms for retaining GTs are the double displacement and the S_Ni mechanisms.

The double displacement mechanism was supported by the first crystal structure of a GT6 member. At a modest resolution of 2.5 Å, the O of the potential nucleophile Glu317 was 3.8-4 Å from C1 of the Gal moiety of UDP-Gal. Although not within covalent bond distance, this could be sufficiently close when the enzyme changed its conformation in the substrate bound state (Gastinel *et al.*, 2001). Following this theory, Patenaude also suggested that Glu303 of GTA and GTB could be a candidate nucleophile when he solved the structures of these enzymes with and without their substrates (Patenaude *et al.*, 2002).

However, this proposal was not accepted because the limited resolutions of the structures of bovine α3GT exhibited insufficient electron density at the link between Glu317 and the Gal moiety. In addition, in the higher resolution structures, there was no electron density observed between the sugar moiety of UDP-Gal and Glu317, and the mutant Glu317Gln activity was reduced but not sufficiently to support the theory

that Glu317 was the catalytic nucleophile (Zhang *et al.*, 2003). Glu317 is important for catalysis by bovine α3GT, reflected in the 30000-fold reduction in catalytic efficiency for Gal transfer to lactose (230-fold for transfer to water) in the Gln mutant and is positioned to act in stabilising a transition state in which the galactose has oxocarbenium ion character in a single displacement S_Ni mechanism, suggesting that the enzyme can follow the S_Ni mechanism (Boix *et al.*, 2002, Zhang *et al.*, 2003).

Nonetheless, the question of the retaining GT mechanism and the GT6 family in particular was raised again when the cavity of the inactive mutant Glu317Ala of bovine α3GT was successfully rescued by azide which acted as a nucleophile to give β-D-galactosylazide (Monegal and Planas, 2006). This finding supported the double displacement mechanism although it did not exclude the alternative S_Ni mechanism. In addition, glycosyl-enzyme intermediates for mutants of GTA and GTB detected by mass spectrometry were reported recently (Soya *et al.*, 2011). In an attempt to test the possibility of Glu303 as a catalytic nucleophile in GTA and GTB activity, the mutant Glu303Cys of both GTA and GTB was made to increase its nucleophilicity. As expected, a peak corresponding to the weight of glycosyl-enzyme was detected. In addition, when the glycosyl-enzyme was incubated with a disaccharide acceptor, a trisaccharide product was detected, suggesting these mutants can catalyse the transfer following the double displacement mechanism (Soya *et al.*, 2011).

Since the recent results were obtained with the modified GT6 enzymes were not definitive evidence for either mechanism, the debate remains open.

1.2.3 Glycosyltransferase from *Bacteroides ovatus*

Bacteroides ovatus is one of the commensal intestinal microbes found in the human gut that is responsible for the pathology of inflammatory bowel disease in humans (Saitoh *et al.*, 2002). This gram-negative bacterium has two genes encoding GT6 glycosyltransferases BoGT6a and BoGT6b, namely ZP_02064961.1 and ZP_02066673.1 respectively. The activity and structure of the former have been characterised, while the latter has only been studied at the sequence level. Thus, knowledge of glycosyltransferases from *Bacteroides ovatus* is from BoGT6a.

1.2.3.1 Characteristic features

BoGT6a (EC 2.4.1.40) is a 263 aa protein containing a N-terminal catalytic domain and a C-terminal membrane-associated, hydrophobic domain (Brew *et al.*, 2010). This enzyme has the same donor substrate specificity as GTA which catalyses the stereochemistry-retaining transfer of GalNAc from the donor UDP-GalNAc to the acceptor 2'-fucosyllactose (FAL) or its homologues (Figure 14). Like its mammalian homologues, BoGT6a is also able to catalyse the hydrolysis of its donor substrate UDP-GalNAc with a higher rate than that of glycosyltransferase (Tumbale and Brew, 2009).

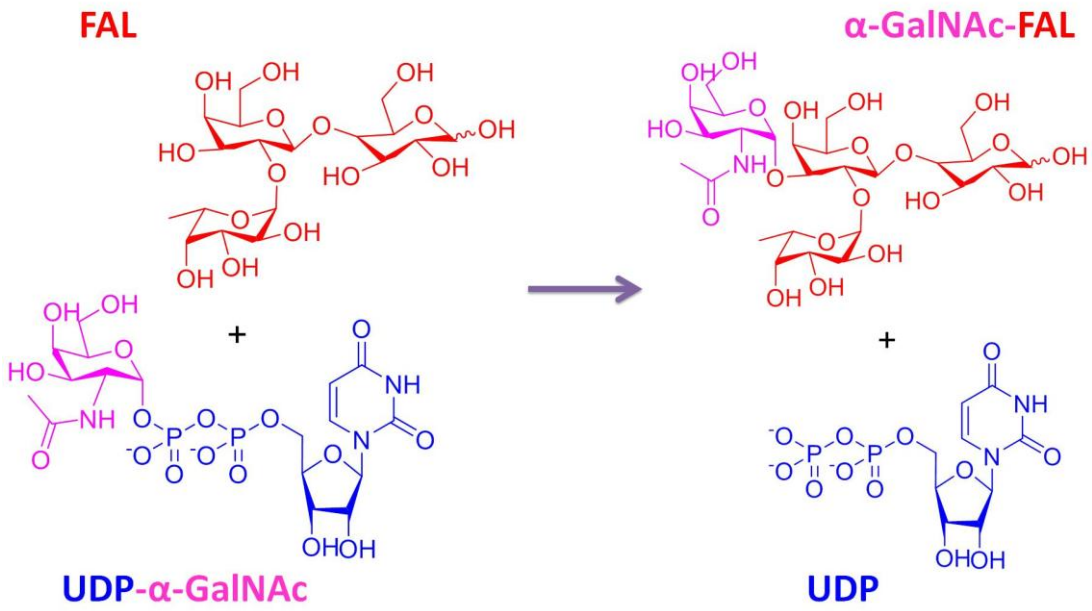


Figure 14. Chemical diagram illustrating the GalNAc transfer from UDP-GalNAc to FAL catalysed by BoGT6a. Diagram was created by using Chemdraw.

In the sequence of the enzyme, the DXD motif which is conserved among GT6 vertebrate members and supposed to interact directly with the metal ion is replaced by a NXN sequence (Asn95Ala96Asn97) (Figure 15). The NXN sequence is also observed in other bacterial GT6 member (Figure 16).

DNA sequence

atg aga att ggt ata tta tat atc tgt act ggc aaa tat gac att ttt tgg aaa gac ttt tat cta agc gca gaa cgt tat ttt atg caa gac caa tct ttc att atc gag tat tat gta ttt act gat agt cct aaa cta tat gac gaa gaa aac aac aaa cat att cac cgg atc aaa caa aag aat tta gga tgg cct gac aac aca tta aaa cgt ttc cat ata ttc ctt cgt atc aag gaa cag tta gag cga gaa acc gac tat cta ttt ttc ttc **aat gcc aat** ctc tta ttc acc agt cct att ggc aaa gaa att cta cca cca tca gat agt aac gga tta cta gga act atg cac cct gga ttc tac aat aaa ccg aac tcc gaa ttt aca tac gag cga aga gat gct tct act gcc tat atc cca gag gga gaa ggt cga tat tat tac gct gga ggg ctt tca ggt gga tgt aca aag gcc tac ttg aaa ctc tgc aca aca att tgc tca tgg gtt gac aga gat gcc aca aac cat ata ata cca att tgg cac gac gaa tct cta atc aat aaa tac ttt tta gat aat cca cca gct att aca ttg tcc cct gca tat cta tac cca gaa ggt tgg ctc ctt cct ttt gaa cca ata atc ctc att cga gac aaa aat aaa ccc caa tat ggc ggg cat gaa tta ttg cga aga aaa aac tct tta tgg gaa agg att aag cta atc tgc caa aaa ttt aaa tcg gct gat **tag**

Protein sequence

M R I G I L Y I C T G K Y D I F W K D F Y L S A E R Y F M Q D Q S F I I E Y Y V F T D S P K L Y D E E N N K H I H R I K Q K N L G W P D N T L K R F H I F L R I K E Q L E R E T D Y L F F F **N A N** L L F T S P I G K E I L P P S D S N G L L G T M H P G F Y N K P N S E F T Y E R R D A S T A Y I P E G E G R Y Y Y A G G L S G G C T K A Y L K L C T T I C S W V D R D A T N H I I P I W H D E S L I N K Y F L D N P P A I T L S P A Y L Y P E G W L L P F E P I I L I R D K N K P Q Y G G H E L L R R K N S L W E R I K L I C Q K F K S A D **Stop**

Figure 15. DNA sequence and Protein sequence of BoGT6a. The 17 C-terminal residues are highlighted in cyan and the NAN in pink.

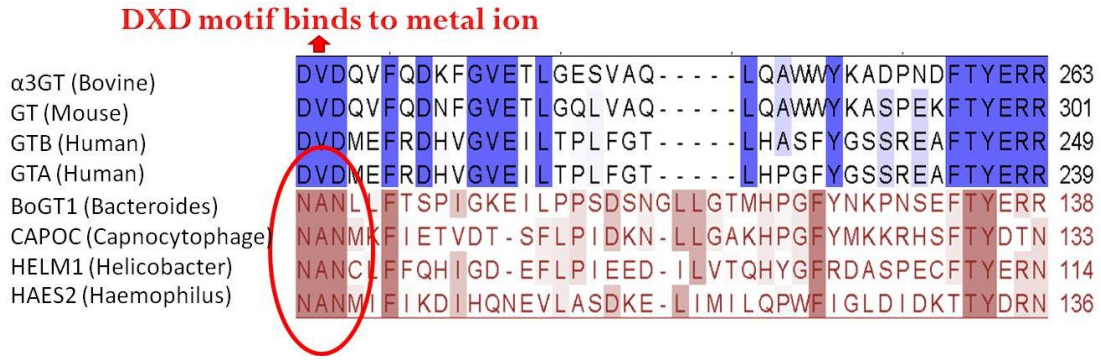


Figure 16. Sequence alignment of GT6 members from both vertebrates and bacteria. The substitution of the DXD motif by the NXN motif is marked. Picture created using ClustalX (Larkin *et al.*, 2007).

As expected, the structurally conservative substitutions of Asn95 and Asn97, corresponding to Asp225 and Asp227 of bovine α3GT, with Asp indicated that the Asn95Asp mutation had a large effect on catalytic activity (more than 4000-fold reduction in k_{cat} and 30-fold increase in the K_M for UDP-GalNAc) much greater than the corresponding substitution for Asn97 (about 8-fold reduction in k_{cat} and 10-fold increase in K_M for FAL). This finding suggested that this Asn95Asp substitution perturbed the interaction of the enzyme with the donor substrate while Asn97 was not involved in donor binding. However, structural studies are needed to determine the role of this region in donor substrate binding (Tumbale and Brew, 2009).

The residue Glu192 of BoGT6a corresponds to the key residue Glu317 of bovine α3GT and Glu303 of GTA/GTB, both of which have considerable importance in catalytic activity of GT6 enzymes (Patenaude *et al.*, 2002, Zhang *et al.*, 2003). Mutagenesis of the corresponding residue of BoGT6a, Glu192, to Gln results in a 10-fold greater reduction of k_{cat} (more than 22,000-fold reduction) than that produced by the same substitution for the corresponding residue, Glu317, of α3GT (Zhang *et al.*, 2003), indicating that Glu192 is a key residue in catalysis in BoGT6a.

Although the sequence of the C-terminal region of BoGT6a differs significantly from those of mammalian GT6 members, alanine scanning indicated that Arg229, Lys231, Arg243 and Arg244 were well conserved among bacterial GT6s and belong to LBR-H and LBR-I which are involved in the C-terminal conformational change induced by the substrate binding. Mutagenesis of these residues indicated that the

Lys231Ala mutation had the greatest adverse effect on the enzyme activity (more than 200-fold k_{cat} reduction compared to 3-fold reduction for Arg229Ala and 10-fold for Arg243Ala/Arg244Ala), which was similar to the effect of the Lys359Ala mutation for bovine α3GT (Jamaluddin *et al.*, 2007, Tumbale and Brew, 2009). The Arg243 corresponds to Arg365 of bovine α3GT, but the insignificant effect on the BoGT6a suggested its role in the cavity of the enzyme may not be as necessary as the role of Arg365 in bovine α3GT activity.

Finally Ala155 of BoGT6a corresponding to Leu/Met266 in GTA/GTB or His280 in bovine α3GT was expected to be involved in the donor recognition specificity. Kinetic assays of the mutant Ala155Met indicated that the enzyme affinity to both UDP-GalNAc and UDP-Gal was reduced about 8-fold and 3-fold respectively, but the catalytic rate of the enzyme with UDP-Gal was increased 5-fold while the enzyme almost lost its activity with UDP-GalNAc (more than 400-fold reduction) (Tumbale and Brew, 2009). This result supports the importance of Ala155 in the donor substrate specificity of BoGT6a.

Based on the knowledge of the GT6 family, the mutagenesis of key residues in BoGT6a that correspond to those that have key roles in catalysis and substrate binding in its mammalian homologues proved that BoGT6a has the same structure-function relationships as mammalian GT6s do. However, BoGT6a retained its activity in the presence of EDTA, which completely inhibited activity of bovine α3GT which requires Mn^{2+}. Increasing the concentration of Mn^{2+} did not enhance the enzyme catalysis, proving that a metal ion is not required for its catalytic activity (Tumbale and Brew, 2009). Sequence alignment showed that the DXD motif, which is well conserved in GT-A glycosyltransferase, was replaced by NXN in bacterial glycosyltransferases (Figure 16). This raised a question about the impact of the NXN substitution on BoGT6a metal-independent activity as well as the catalytic mechanism of BoGT6a in particular and GT6 family members in general.

1.2.3.2 Native structure

The active domain (1-246) of BoGT6a was expressed with an N-terminal His-tag. The protein was expressed by *E. coli* BL21 (DE3) cells and purified by affinity

chromatography with Ni^{2+}-NTA Super Resin (Qiagen) (Tumbale and Brew, 2009). The structure of BoGT6a in substrate free form was solved by MR at 3 Å resolution, consisting of 10 β-strands, 4 α-helices and three 3_{10} helices following the GT-A fold with a Rossmann-like domain (Thiyagarajan *et al.*, 2012) (Figure 17). The only remarkable structural difference is its short N-terminal region which supposedly relates to localisation of mammalian glycosyltransferase in the Golgi (Tu and Banfield, 2010) (Figure 18).

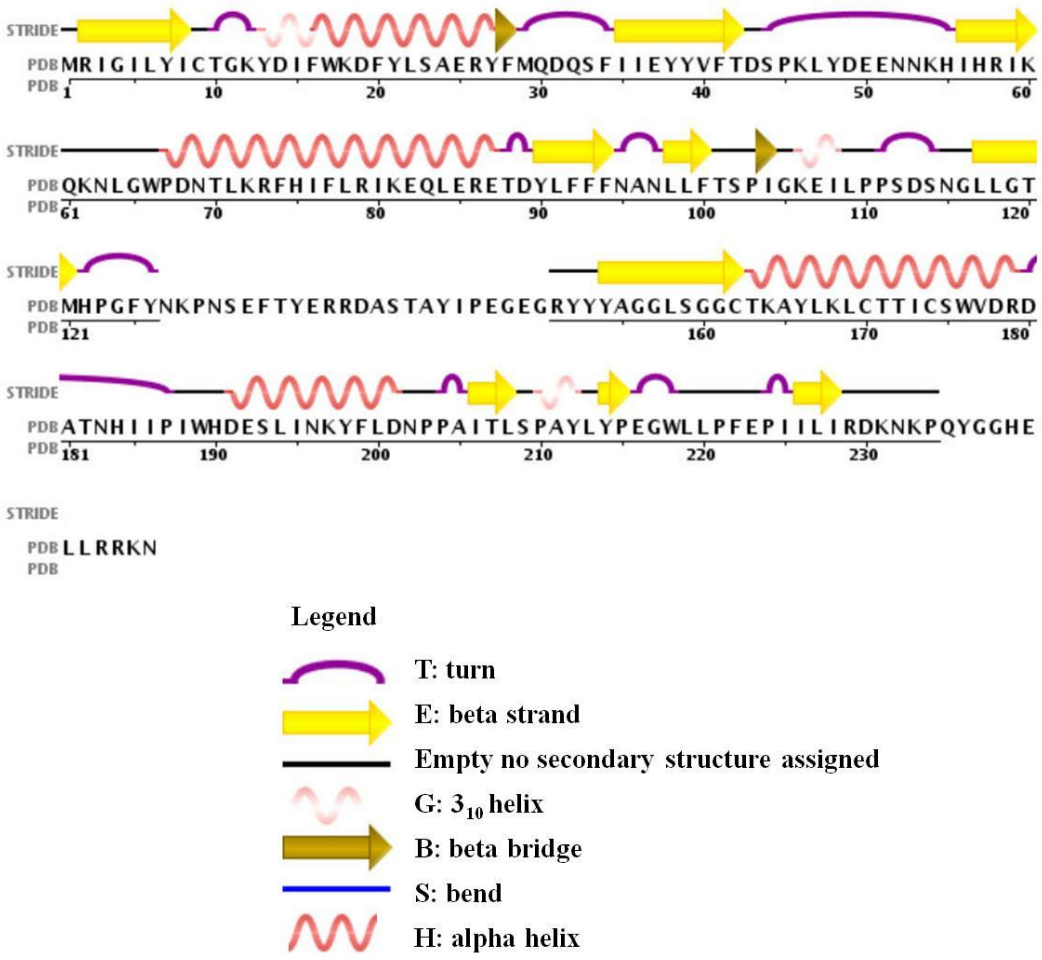


Figure 17. Secondary structure of BoGT6a apo form. The image was sourced from RCSB source (Velankar *et al.*, 2005, Golovin *et al.*, 2005, Kabsch and Sander, 1983).

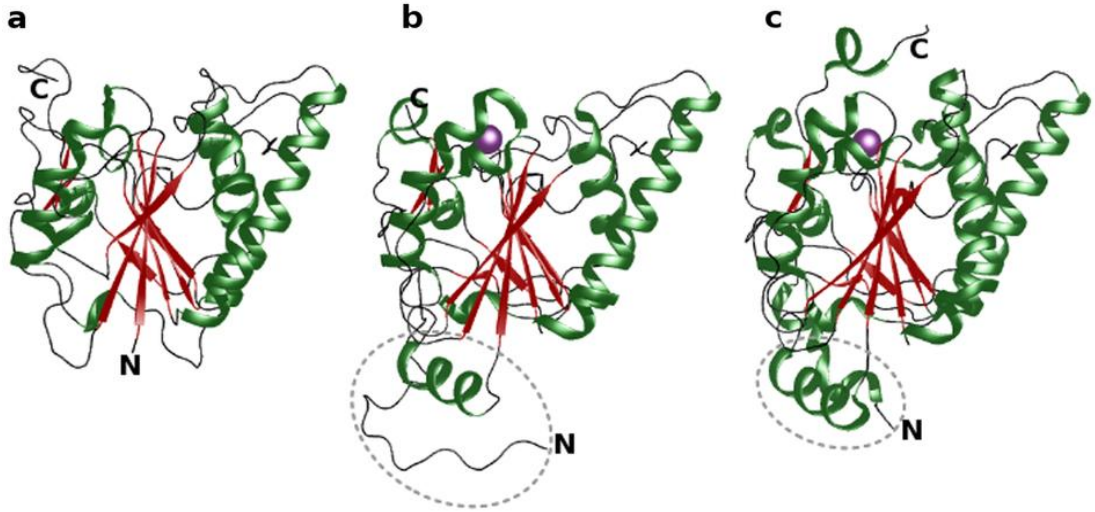


Figure 18. N- and C- terminus of BoGT6a apo form (PDB: 4AYL) (a), GTA (PDB: 1ZI1) (b), and bovine α3GT (PDB: 1GX4) (c). The N-terminal extension in GTA and α3GT are marked. The manganese ions in GTA and α3GT are shown as purple spheres (Thiyagarajan *et al.*, 2012).

As is the case for its mammalian homologues, BoGT6a also contains 2 flexible loops, namely an internal mobile loop (from residue 126 to residue 151) and a flexible C-terminus (from residue 234 to residue 246) which were disordered in its apo form structure. These regions are absent in the apo form structure, resulting in a structure consisting of two regions, namely region 1 from Met1 - Y126 and region 2 from Arg151 - Pro234 (Figure 17).

There are only two ions presenting in the structure, Cl^- located near a loop between Met 29 and Phe 34, and Ca^{2+} interacting with C-terminal residues Glu 216, Asp 230 and Asn 232; there is no metal ion observed in the active site of BoGT6a (Figure 18). This is consistent with its metal-independent activity (Thiyagarajan *et al.*, 2012).

A superposition of the structure of BoGT6a with the structure of bovine α3GT (PDB 1K4V) showed positions of 9 conserved LBRs observed in previous published structures of GT6 family; LBR-A includes residues 7-13, LBR-B residues 70-72, LBR-C residues 95-97, LBR-D residues 122-125, LBR-E residues 153-157, LBR-F residues 189-192, LBR-G residues 215-218, LBR-H residues 231-236, LBR-I residues 243-245 (Figure 19).

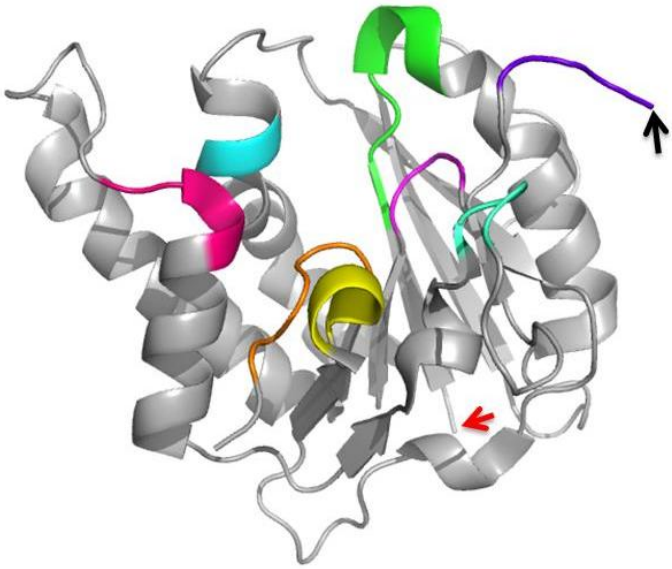


Figure 19. Positions of nine conserved LBRs on the BoGT6a apo form structure (PDB 4AYL). LBR-A is coloured in green, LBR-B in cyan, LBR-C in magenta, LBR-D in yellow, LBR-E in orange, LBR-F in hot pink, LBR-G in green cyan, and LBR-H in purple blue. The protein is shown in cartoon representation. The black arrow points out the C-terminus and the red arrow the N-terminus. Picture created using Pymol.

The striking structural similarity to GT6 enzymes led to the proposal that the difference in the activity of GT6 members with respect to metal ions could be caused only by the change of individual residues in the active site. Investigating this requires structural studies of BoGT6a in complex with its donor substrate, UDP-GalNAc, and its acceptor substrate, FAL, which will provide an insight into how key residues interact with the ligands and their roles in the enzyme activity.

1.2.3.3 Catalytic mechanism

The catalytic mechanism of BoGT6a is not well-understood just like all vertebrate GT6 members. This is true even with the two proposed mechanisms: the S_Ni mechanism and the double displacement mechanism.

In the S_Ni mechanism BoGT6a is said to play a role as a coordinator, orientating the donor substrate, UDP-GalNAc and the acceptor substrate (FAL), bringing them close to each other (Figure 20A). At an approximate distance the lone pair on the oxygen atom of FAL attacks the C1 of the GalNAc moiety of UDP-GalNAc, breaking its

bond with UDP moiety and forming a bond between C1 of GalNAc and O3 of FAL. The configuration of GalNAc remains as the α conformation without the presence of the glycosyl enzyme intermediate.

In the double displacement mechanism the bond between C1 of GalNAc moiety and UDP moiety is broken by an attack from the lone pair on the oxygen atom of the hydroxyl group of Glu192. This attack leads to a formation of glycosyl-enzyme intermediate and the conformation of GalNAc moiety changes to the β conformation. The glycosyl-enzyme intermediate undergoes a nucleophilic attack by the oxygen atom of the FAL acceptor substrate forming a bond between C1 of GalNAc and O3 of FAL to give the final product. This second nucleophilic attack changes the conformation of GalNAc back to α conformation (Figure 20B).

Currently, there has been only one successful attempt at trapping a glycosyl-enzyme intermediate of GTA and GTB mutant Glu303Cys (Soya *et al.*, 2011). There is no structural evidence of a glycosyl-enzyme intermediate reported. One of the reasons may be that the native GT6 members catalyse their reaction at a high rate meaning that, the intermediate is not stable enough to be observed by crystallography. A structural study of both the native form and the mutant form of BoGT6a (with reduced activity) in a complex with its donor and acceptor substrate may illustrate the mechanism of BoGT6a in particular. This may be applied to other GT6 members.

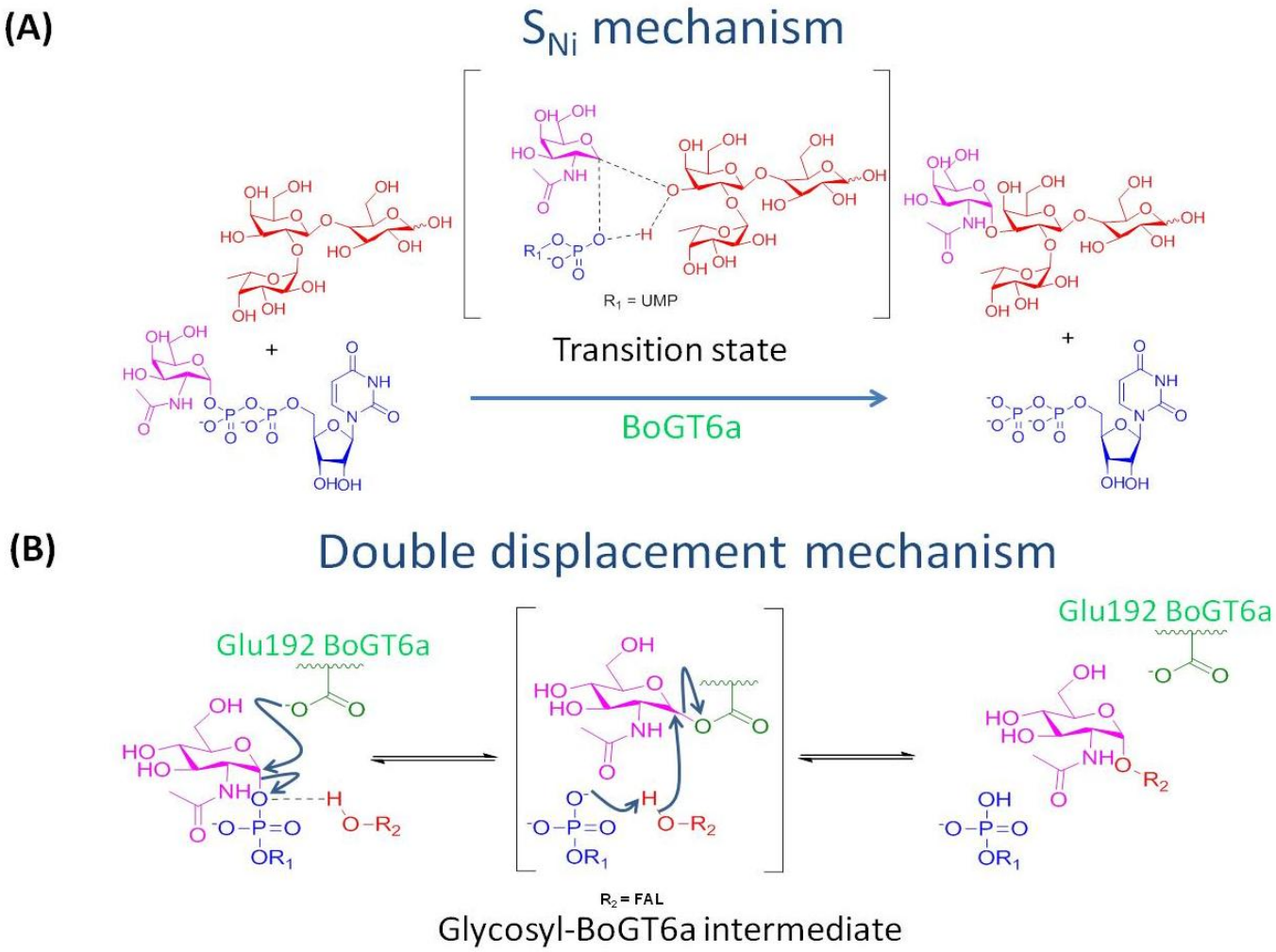


Figure 20. Proposed mechanisms for BoGT6a catalytic activity. (A) the S_Ni mechanism. (B) the double displacement mechanism. R1 is noted for UMP and R2 for the acceptor FAL. The diagram was created using ChemDraw.

1.3 Aims and objectives

1.3.1 Aims

The main goal of this project was to elucidate the possible role of the substitution of Asp-X-Asp motif in the metal- independent catalytic mechanism of BoGT6a. In particular, structural studies of BoGT6a with its potential donor UDP-GalNAc and/or acceptor FAL were undertaken to determine whether the mutated residues perturbed the interaction between enzyme and donor substrate, or if mutated residues changed the active site structure of the enzyme. In addition, structural characterisation of the BoGT6a Glu192Gln (BoGT6a E192Q) mutation was performed to analyse the impacts of key residues on the catalytic activity of the enzyme itself, and of the GT6 family in general.

1.3.2 Objectives

1. Determine the structure of native BoGT6a in complex with the donor (UDP-GalNAc) and/or acceptor (2'-fucosyllactose).

2. Determine the structure of BoGT6a E192Q, which showed a significant reduction in catalytic activity, in complex with the donor (UDP-GalNAc) and/or acceptor (2'-fucosyllactose).

CHAPTER II

Structure of BoGT6a in complex with 2’-fucosyllactose

2 Structure of BoGT6a in complex with its acceptor substrate 2'-fucosyllactose

2.1 Methods

The truncated form of BoGT6a, which lacked 17 residues at the C-terminal end and terminated at Asn246, was expressed and purified as described by Tumbale *et al.* (2009). This C-terminal truncated form has been shown to be fully active and was provided to us by our collaborator Professor Keith Brew, Florida Altantic University, USA.. This form is referred to as BoGT6a native protein throughout the thesis and the mutants are derived from it.

Crystals and diffraction data of BoGT6a in complex with FAL were obtained by our former colleague Dr. Amit Sundriyal. The crystal of BoGT6a·FAL diffracted to 2.67 Å resolution and 200 images were collected at Diamond Light Source, but the resolution was cut off at 3 Å such that the R_{merge} is acceptable. The dataset was indexed using HKL2000 (Otwinowski and Minor, 1997) in space group $P2_1$. The mtz file was derived from the scale file obtained from HKL2000 by using the import scale file function in CCP4i (Winn *et al.*, 2011).

Phaser version 2.3.0 (McCoy *et al.*, 2007) in CCP4i version 6.2.0 (Winn *et al.*, 2011) was used to calculate phases for the BoGT6a•FAL complex data. The MR method was used (using the fully refined native BoGT6a structure as the starting model) in space group $P2_1$ with four molecules in the asymmetric unit. The missing loop (from residue 127 to residue 150) in the native BoGT6a structure was built based on its clear electron density. Clear electron density was observed for the acceptor substrate FAL and the ligand was inserted into the structure. Further refinement and model building were carried out using PHENIX.refine version 1.7.1_743 (Adams *et al.*, 2010) and COOT (Emsley and Cowtan, 2004) (for model building). 5% of reflections were used to calculate R_{free}, and refinement and model building continued until the R/R_{free} factors reached 18.07/26.19 and acceptable Ramachandran values were achieved. In spite of the moderate resolution, the structure is highly ordered and the information about the structural changes due to acceptor molecule recognition is quite evident in all four copies of the protein in the asymmetric unit.

2.2 Results

The crystal belonged to the monoclinic space group $P2_1$. The Matthew's coefficient and solvent content were calculated from the unit cell dimensions and the molecular weight of BoGT6a using Matthews_Coef (Matthews, 1968, Winn *et al.*, 2011, Kantardjieff and Rupp, 2003). Solvent content was calculated to be 47% with a 0.7 probability across all resolution ranges (*P(tot)*) with 4 molecules per asymmetric unit. The phase problem was solved by MR method using the Phaser with the native BoGT6a structure as the search model. The resulting structure had 4 chains with LLG and TFZ values of 2217 and 38.2 respectively. Each chain contained only the first 230 residues of BoGT6a. The missing region from 127 to 150 which was not observed in the native BoGT6a due to its disorder was clearly visible in the electron density map of the complex BoGT6a•FAL (Figure 21).

After a number of cycles of refinement, clear difference density was visible for both the flexible region at the C-terminus and ligand of each chain, and so these were built into the structure (Figure 22 and Figure 23). Refinement was performed to improve quality of the structure until R and R_{free} got 18.07 and 26.19 respectively. 96.7 % of residues are in the favoured region of the Ramachandran plot and 3.3 % are allowed (calculated by Procheck program). Other statistics are summarised in Table 2.

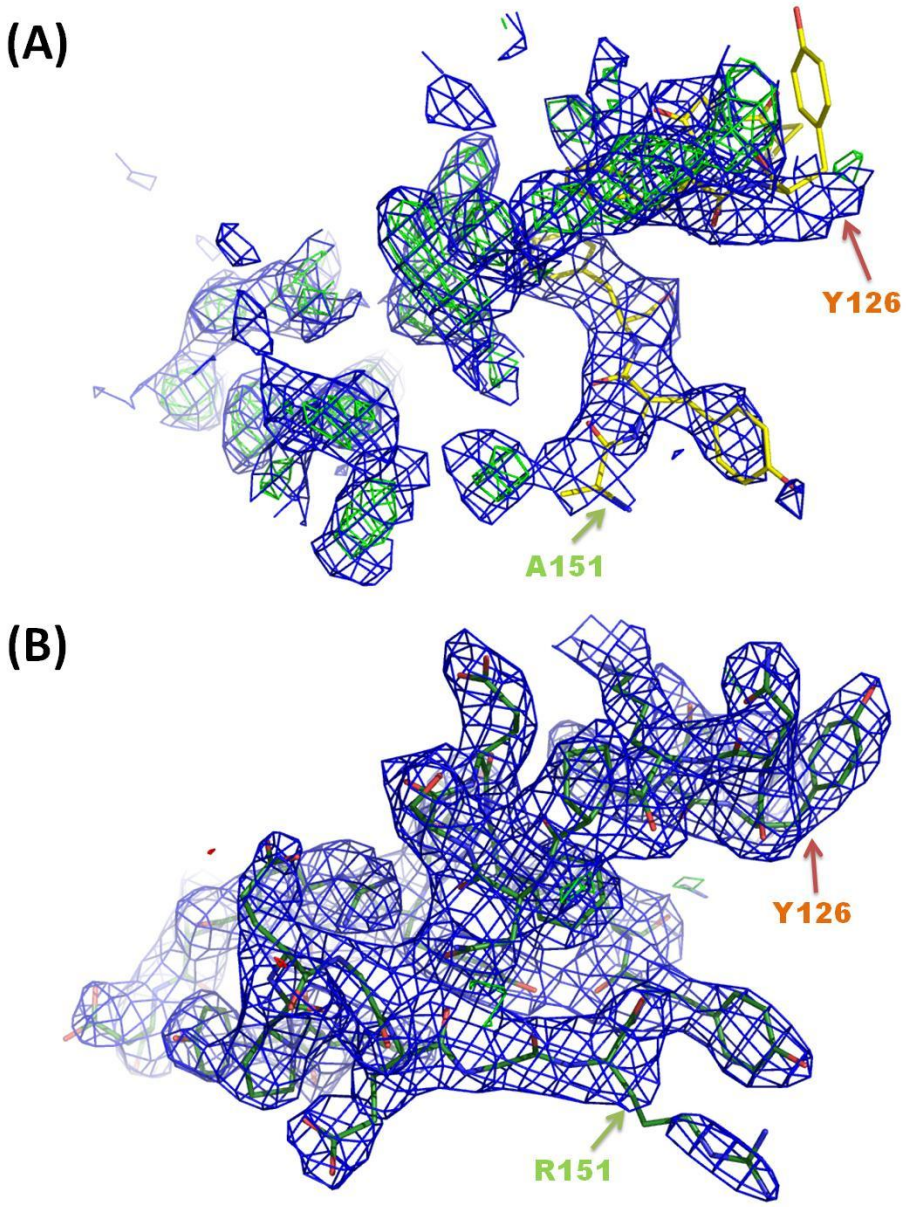


Figure 21. Electron density map observed in BoGT6a•FAL structure (chain B as representative) before and after the flexible loop from residue 126 to 151 was built. (A) BoGT6a•FAL without residues 126 – 151, and (B) BoGT6a•FAL with residues 126 – 151. Protein is shown in yellow for chain B without the loop, and green for chain B with the loop. Residues are noted in 1 letter abbreviation, showing the loop position. The $2F_c$-F_o map is contoured in blue at 1σ, and the F_c-F_o map in green (positive) and red (negative) at 3 σ. Picture created using Pymol.

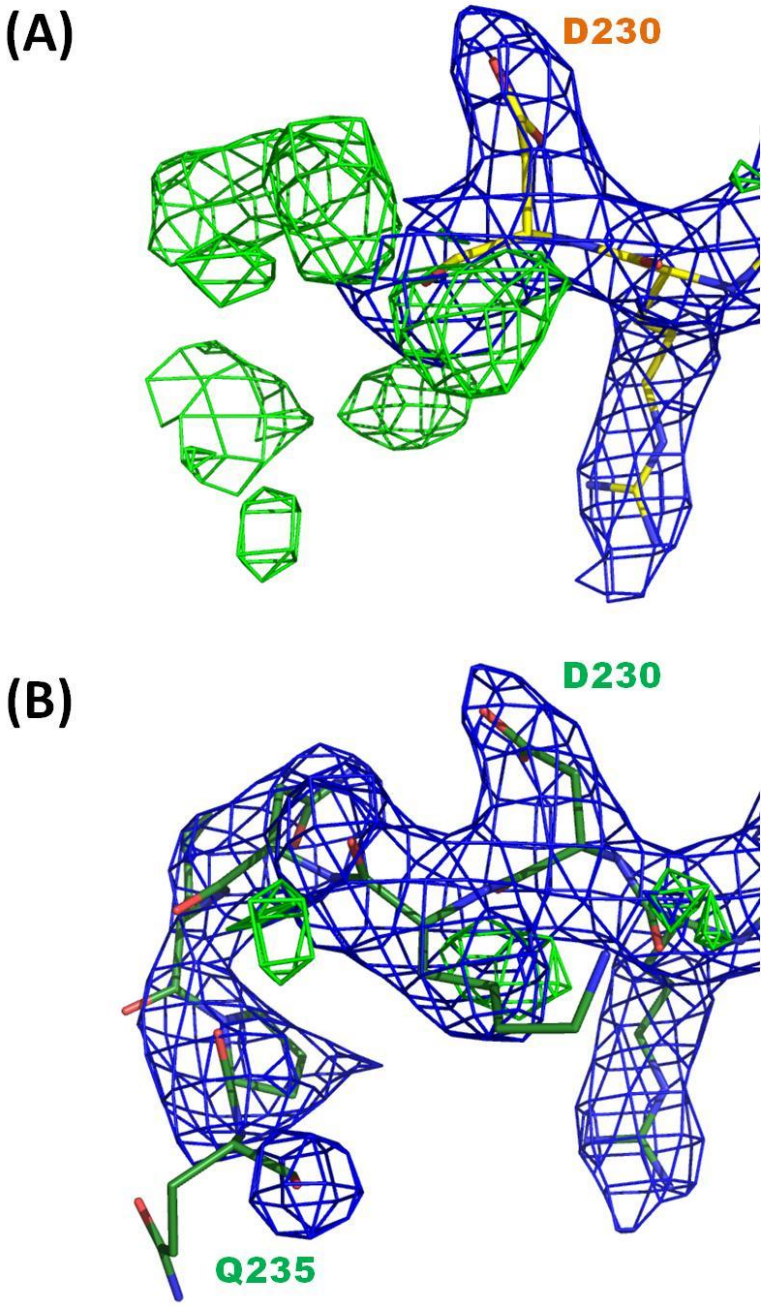


Figure 22. Electron densities of C terminus (of chain B as representative) in the BoGT6a•FAL. (A) the different densities of C terminus after phasing with BoGT6a apo form as starting model. (B) shows the final electron densities fitted well with the end of the C terminus. Protein is shown in yellow for chain B without the end of the C terminus, and green for chain B with the end of the C terminus. Residues are noted in 1 letter abbreviation, showing the C terminus. The $2F_c$-F_o map is contoured in blue at 1σ, and the F_c-F_o map in green (positive) and red (negative) at 3 σ. Picture created using Pymol.

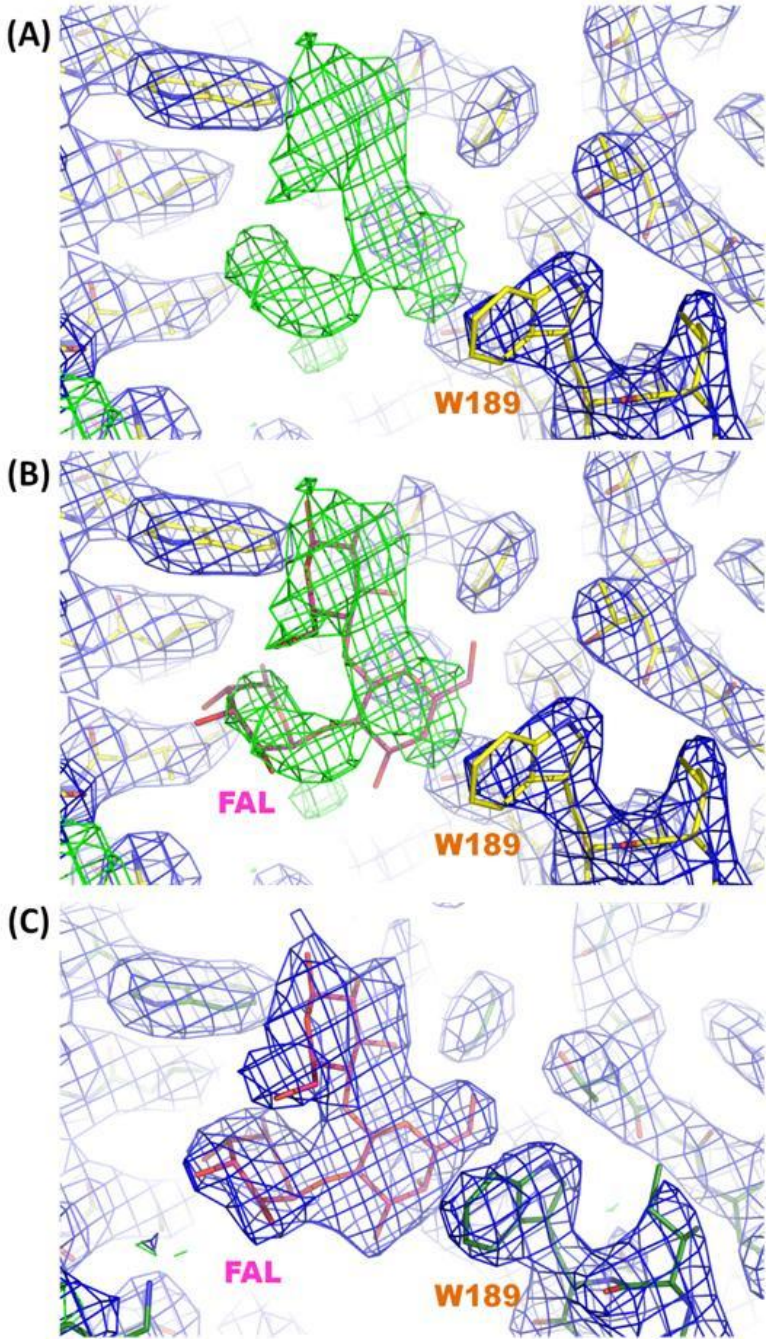


Figure 23. Electron densities of FAL (of chain B as representative) in the BoGT6a•FAL structure. (A) the difference electron density of FAL at the active site after phasing with BoGT6a apo form as starting model. (B) shows how FAL moiety was supposed into the difference electron density. (C) shows the final electron density fitted well with the FAL moiety. Protein is shown in yellow for chain B without FAL, and green for chain B with FAL. FAL is shown in pink. Residue Trp189 is noted in 1 letter abbreviation, showing the active site of BoGT6a. The $2F_c$-F_o map is contoured in blue at 1σ, and the F_c-F_o map in green (positive) and red (negative) at 3 σ. Picture created using Pymol.

Table 2. Data collection and refinement results for BoGT6a•FAL structure

Space group	Monoclinic, P 2_1
Number of molecules per asymmetric unit	4
Cell dimensions	a= 70.9 Å, b=93.9 Å, c=75.5 Å, β=93.8^o
Resolution range (Å)	70.7 – 3.0 (3.2 – 3.0)
$R_{p.i.m.}$ (outer shell)	0.07 (0.29)
$I/\sigma I$ (outer shell)	9.3 (2.6)
Completeness (outer shell) %	94.1 (93.7)
Total no. of reflections	72365
Unique no. of reflections	18641
Redundancy (outer shell)	3.9 (3.8)
Wilson B-factor (Å^2)	47.78
R_{cryst}/R_{free} (%)	18.07/26.19
Average B-factor (Å^2) Overall Protein Ligand (FAL)	 36.6 A: 30.5, B: 35.2, C: 41.7, D: 38.8 36.2
RMSD bond length (Å) bond angle (^{o})	 0.009 1.429
Number of protein atoms Number of ligand atoms	A: 1928, B: 1961, C: 1945, D: 1973 132
Ramachandran plot statistics (%) Favoured Additionally allowed	 96.7 3.3
PDB ID	4AYJ

The final structure consists of 4 molecules in an asymmetric unit (Figure 24). There are 231 residues in chain A, 235 residues in chain B, 233 residues in chain C, and 236 residues in chain D. The topologies of all the four subunits are similar to each other, with an r.m.s.d value of 0.41 Å for chain A and chain B, 0.39 Å for chain A and chain C, and 0.47 Å for chain A and chain D (calculated by using the program COOT).

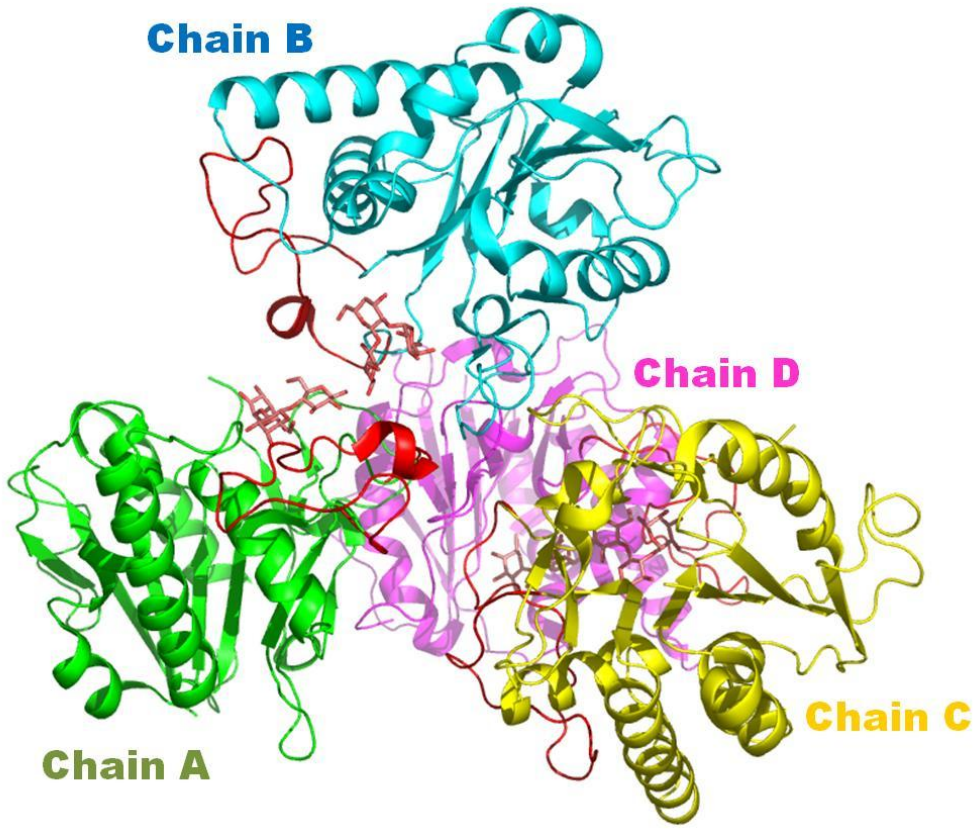


Figure 24. Crystal structure of BoGT6a in complex with FAL. The 4 molecules in an asymmetric unit are coloured by chain. The flexible loop from residue 126 to residue 151 (loop 1) of each chain is coloured in red. Ligands are shown as stick in pink. Picture created using Pymol.

The structure of the complex consists of 10 β-strands, 4 α-helices and 2 3_{10}-helices as calculated by STRIDE (Heinig and Frishman, 2004) (Figure 25). Its general structure is similar to the BoGT6a apo form which follows the GT-A fold and is strikingly similar to those of other enzymes in the GT6 family, with a central β sheet made of 8 β-strands surrounded by 4 α-helices. The structure can be divided into two

domains (Figure 25). The first domain beginning from Met1 to Phe94 is comprised of a central β sheet composed of 3 N-terminal β-strands (β1, β2, and β3) and two surrounding α-helices (α1, and α2). The other domain, from Asn95 to Pro234, consists of a β-sheet composed of 2 anti-parallel β- strands (β4 and β7) and 2 parallel β-strands (β6 and β8), at the centre, two long α-helices (α4 and α5) on one side, and then a pair of anti-parallel β-strands (β5 and β10). The loop region from residue Tyr127 to Gly150 (loop 1), which is absent in the structure of the apo form of BoGT6a due to a lack of visible electron density, can be seen clearly in the structure of the complex as a large loop (Figure 24).

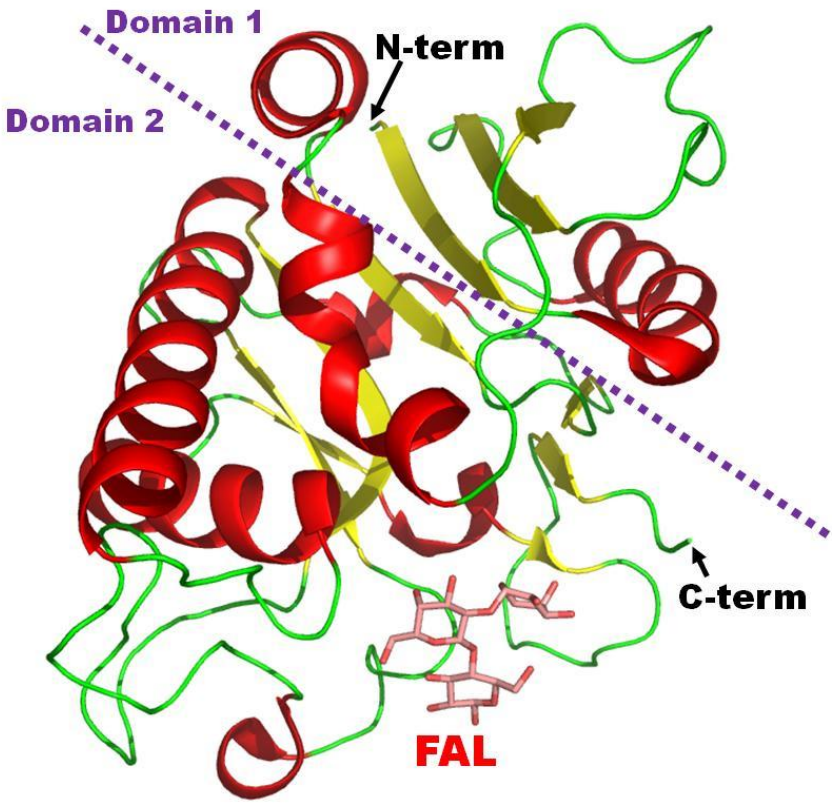


Figure 25. Secondary structure of BoGT6a in complex with FAL. Protein is shown in cartoon representation and coloured by secondary structure. Two domains are distinguished by a purple dash line. N terminus and C terminus are noted. The ligand is shown as stick in pink. Picture created using Pymol.

2.3 Discussion

The conformation of the region from residue Met1 to Tyr126 in the BoGT6a•FAL complex is similar to that of the same region in BoGT6a apo form, with an r.m.s.d value of 0.54 Å. In contrast, the conformation of the region from Ala151 to Lys231 of the BoGT6a•FAL complex is quite different from that of native BoGT6a, with an r.m.s.d value of 1.74 Å (calculated using the program COOT). This region contains the loop from residue 181 to 192 (loop 2) and the C-terminal region from residue 229 to 231 (C-term) (Figure 26). This suggests that the C terminal region, loop 1 and loop 2 could play important roles in BoGT6a catalytic activity. In addition, the interaction between the residues of this region and FAL stabilises the loop conformation, also pointing to its importance in catalytic activity.

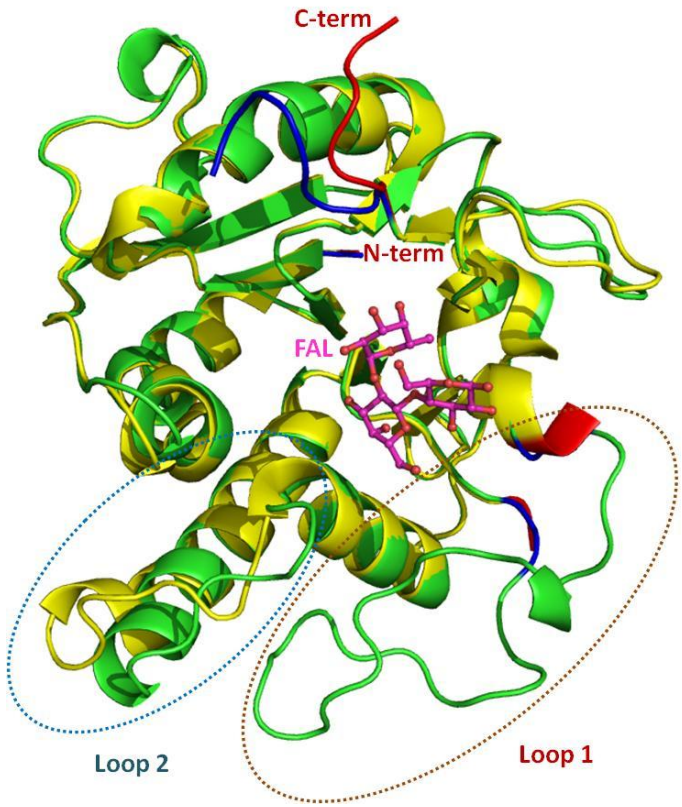


Figure 26. Overall structure comparison between of BoGT6a•FAL compared to BoGT6a in substrate free form. Proteins are shown in cartoon representation in yellow for BoGT6a apo form and in green for BoGT6a•FAL, residue 1-2, 125-126, 151-152 and 226-236 are marked in red for BoGT6a apo form and in dark blue for BoGT6a•FAL. FAL is shown as stick and coloured in magenta. The C-termini and N-termini are labelled. Two flexible loops which were changed significantly are circled. Picture created using Pymol.

Analysing the packing of molecules in the BoGT6a•FAL showed there were two 2-fold axes. The first one is the axis between residues Pro221 of molecule A and molecule B. The second axis is the axis going through between residues Pro221 of chain D and chain C (Figure 27A). Each chain interacts with FAL at the C-terminus region. The packing arrangement of the four chains means that in each pair (A and B, C and D) the acceptor substrate binding sites face each other (Figure 24).

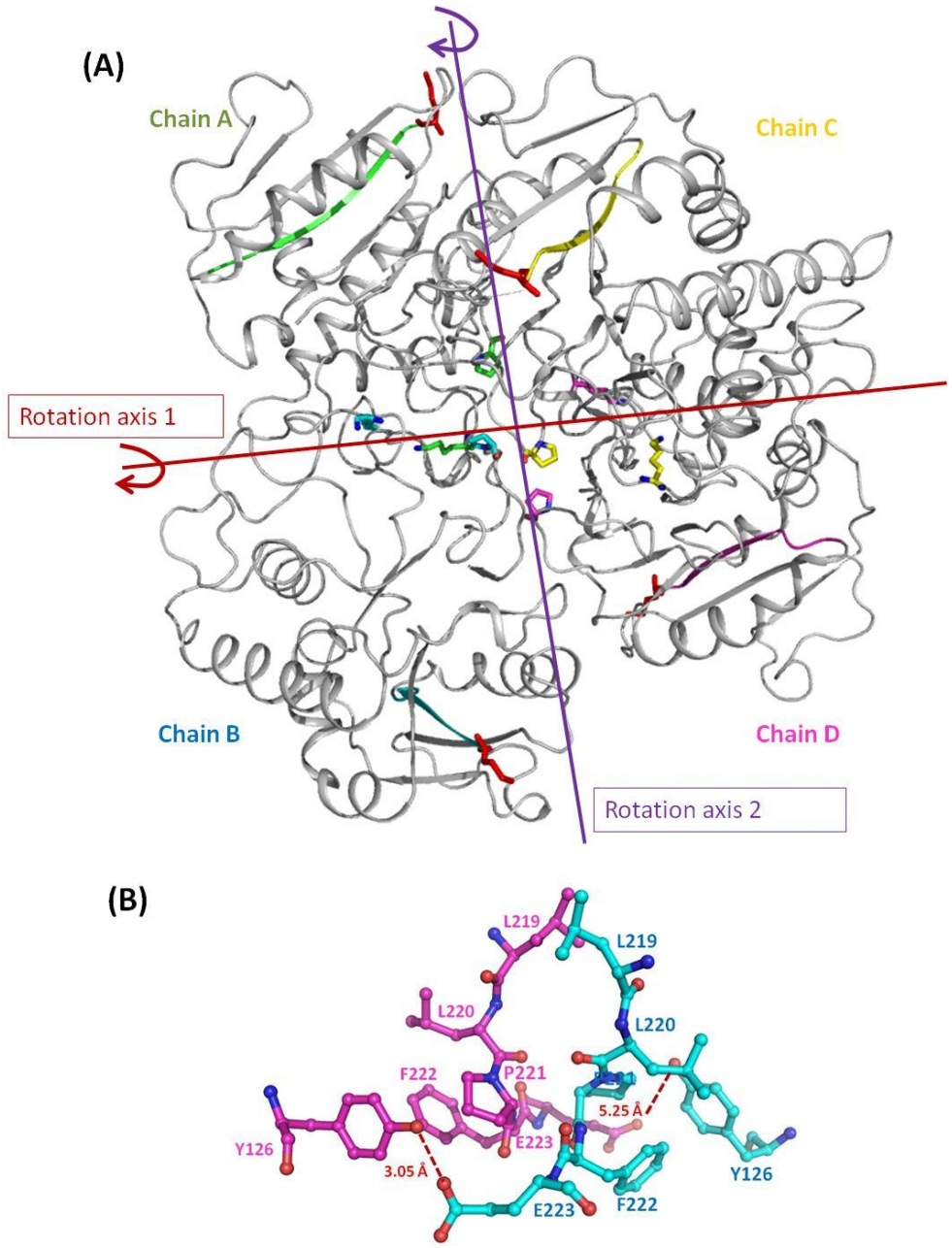


Figure 27. Symmetry in BoGT6a•FAL complex crystal packing. (A) Rotation operation applied to the structure of BoGT6a•FAL. Protein displayed in cartoon representation in

which the first 10 residues of each chain are coloured in green for chain A, cyan for chain B, yellow for chain C and magenta for chain D. Pro221 and Lys128 are shown as sticks and coloured following the chain that they belong to. The first residue, Met1, in all chains is contoured as red sticks. (B) Symmetry between chain B (in cyan) and chain D (in magenta). Protein shown in stick-ball model, and the distances between residues Y126 and E223 displayed as red dashes. The picture was created using Pymol.

There are 4 cysteine residues in BoGT6a sequence. Cys9 belonging to LBR-A is buried in the active site of BoGT6a. The other cysteines, Cys162, Cys170 and Cys174 belong to helix α7 in which only Cys162 and Cys174 are exposed on the protein surface (Figure 28).

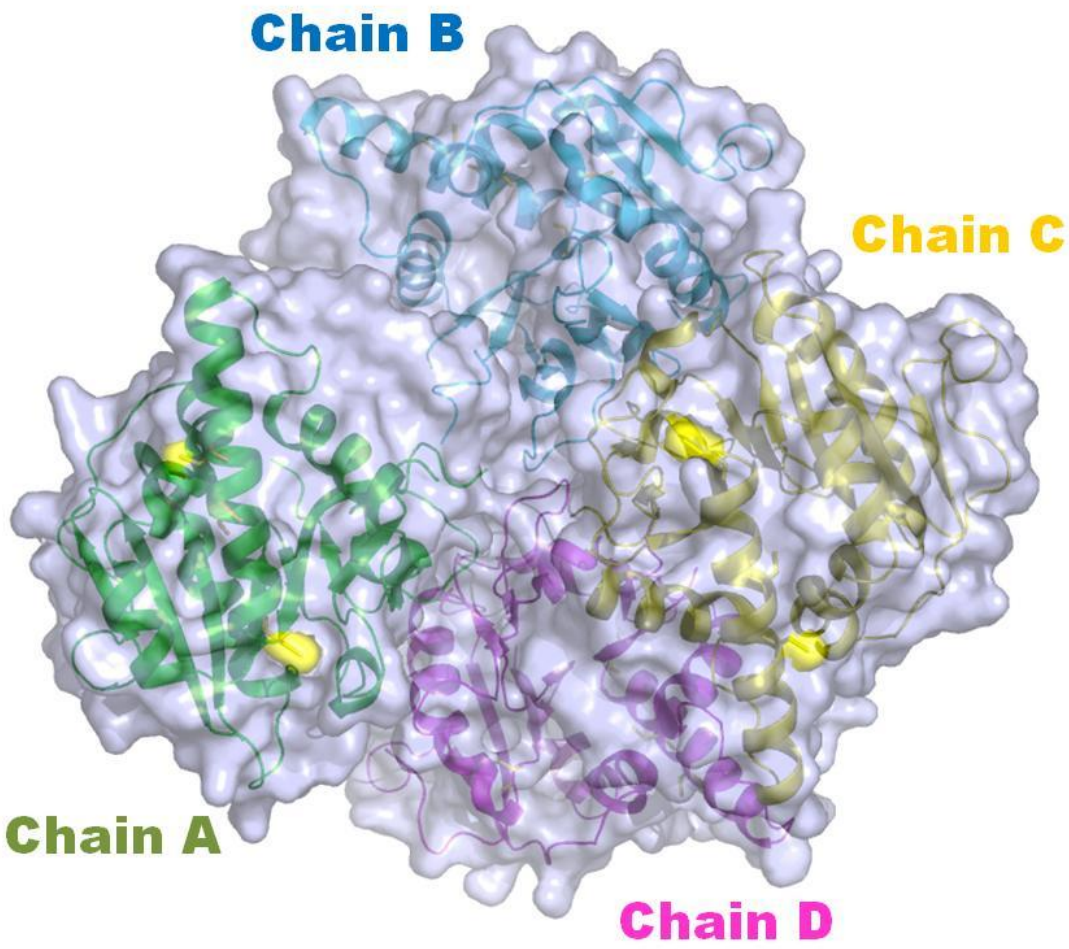


Figure 28. Surface of BoGT6a•FAL structure. Protein is shown in cartoon representation and coloured by chain. Surface of all residues are coloured in light blue, except those of 4 cysteine residues. Cysteine residues are shown as stick and surface in yellow. There are only surfaces of Cys162 and Cys174 exposing on the surface of the protein. FAL moieties are

shown as pink sticks. The others are buried in the core of the protein. Picture created using Pymol.

However, as these exposed Cys belonging to a helix which is high ordered, they may highly be involved in stabilising the alpha helix rather than forming any disulphide bridge between molecules. This agrees with the finding reported by Tumbale *et al.*(2009) that BoGT6a is active in a monomer form (Tumbale and Brew, 2009). In addition, regardless to the appearance as two dimers in an asymmetric unit, buried surface analysis of both the complex structure with and without ligand using PISA (Krissinel and Henrick, 2007) indicates that the dimer interface results from the packing of the crystal and does not pertain to a biological unit. Of the 4 chains, chain A has the lowest B factor value and hence this chain was used as a representative molecule of the BoGT6a•FAL complex for further analysis.

The disordered region between Tyr126 and Gly150 observed in the native BoGT6a structure appears to be ordered when it interacts with FAL. This region forms a large loop (loop 1) which is responsible for acceptor binding (Figure 26). Residues His122, Lys128, and Glu132 from this loop along with Trp189 and Glu192 form hydrogen bonds with FAL (Figure 29). The acceptor binding site also has many hydrophobic residues (Trp189, Trp218, Pro215, Ile228, Pro123, Phe125 and Tyr153) which are important in accommodating and maintaining the correct orientation of the acceptor molecule (Figure 29). Sequence alignments have indicated that Tyr126 of BoGT6a corresponds to residue Tyr172 of GTA and bovine α3GT, and that residue Gly150 corresponded to residue Gly196 of GTA and bovine α3GT (Tumbale and Brew, 2009). This means that the disordered loop in BoGT6a corresponds to the internal flexible loop in its mammalian homologues (residues 176–188 in GTA/GTB, and residues 188 – 199 in bovine α3GT) but it is larger with 24 residues. In addition, loop 2 (residue Ala181 – Glu192) also underwent a conformational change in which Trp189 was brought into the acceptor binding site, suggesting that the loop conformation and its residues could play an important role in the activity of the enzyme (Figure 30). This is expected because residues 189-192 belong to LBR-F which is a region directly interacting with both donor and acceptor substrates (Heissigerova *et al.*, 2003).

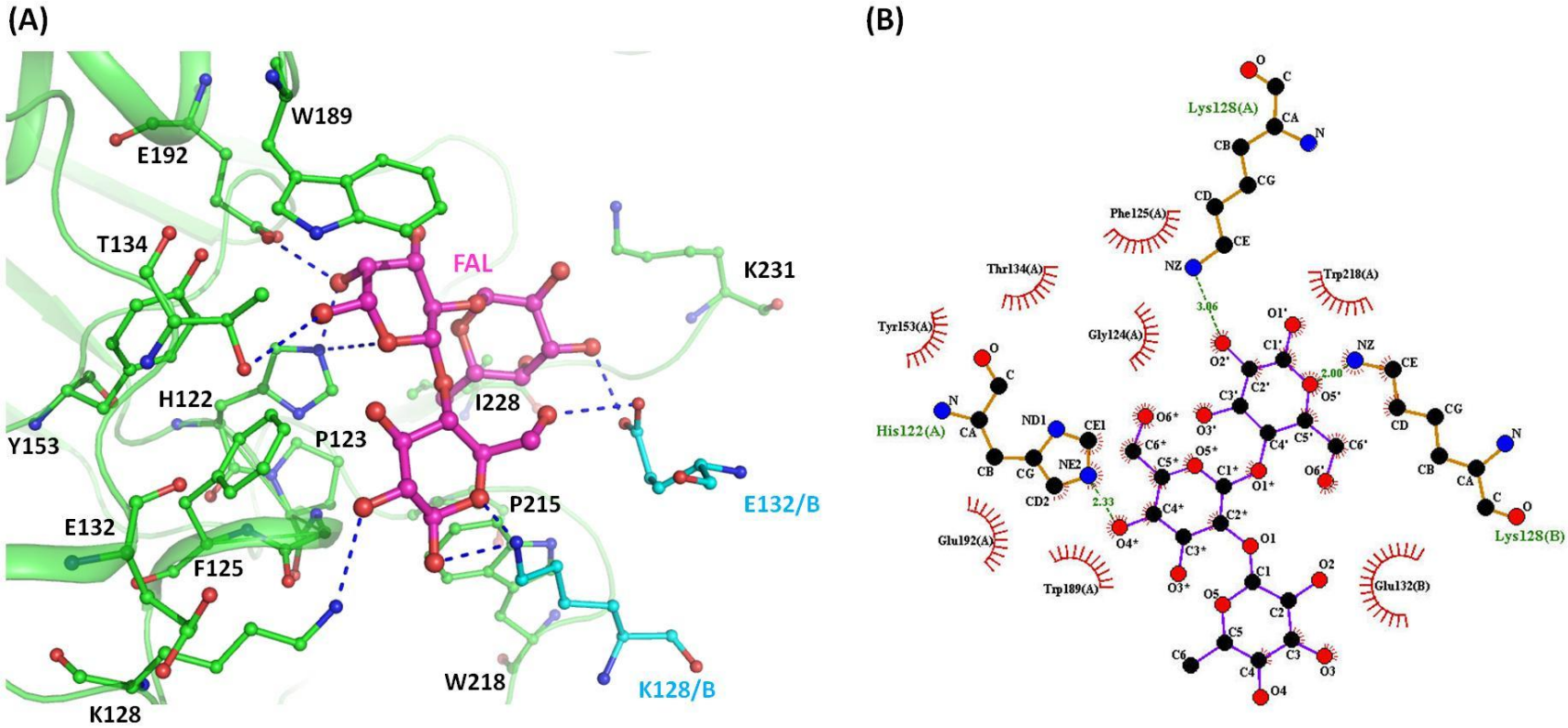


Figure 29. Interactions between BoGT6a and its acceptor substrate, FAL in the acceptor binding site. (A) Interactions of BoGT6a and FAL in the acceptor binding site. The Protein is shown in cartoon representation, and interacting residues and FAL as stick-ball. The interacting residues from chain A are coloured in green, the residues from chain B in cyan, and FAL in magenta. The possible hydrogen bonds are shown as blue dashes. The image was created using Pymol. (B) Ligplot of BoGT6a-FAL (chain A as representative) interactions with key residues. FAL is shown in purple, interacting residues in orange, and hydrophobic interacting residues in red colours. Hydrogen bonds are shown as green dashes. Image created using Ligplot (Wallace *et al.*, 1995).

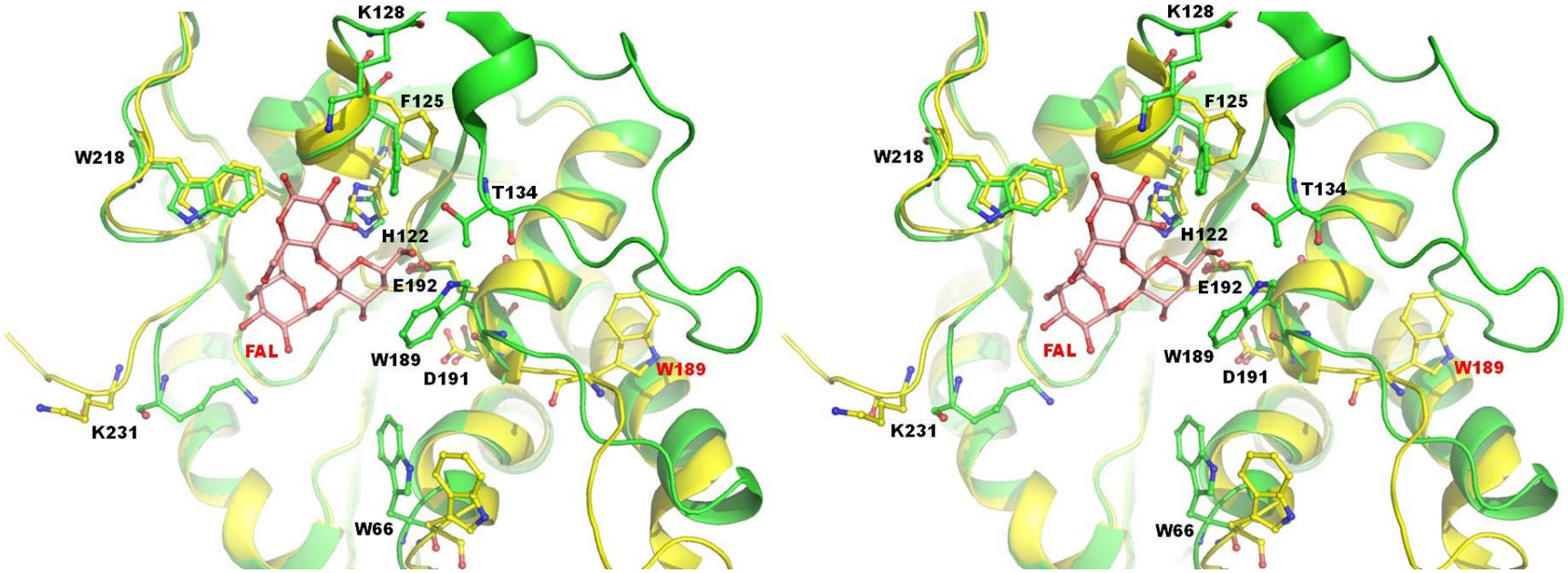


Figure 30. Stereo view showing conformational change of individual residues in the acceptor binding site of BoGT6a in complex with FAL compared to those of BoGT6a in substrate free form. Proteins are shown in cartoon representation in yellow for BoGT6a apo form and in green for BoGT6a•FAL. FAL is shown as stick and coloured in pink. The interacting residues are shown as stick-ball model and coloured following the protein colours as above. The picture was created using Pymol.

The orientations of the C-terminal residues of BoGT6a in complex with FAL are almost completely opposite to those in the BoGT6a apo form. The C-terminal region in the complex structure tends to cover the active site of the enzyme, while in the apo form of the structure the C-terminal region appears more open (Figure 26, Figure 30). This conformation is in agreement with those observed in other GT6 members where substrate binding also induces a conformational change of the C-terminal region. However, strong electron density was only observed for residues up to Lys231. Further residues which were observed in other chains were highly flexible (average B factor values of residues from 231 to the end of C-terminus in chain B, C and D vary from 47 to 72 $Å^2$) and there is a missing region at the C-terminus (from Gly237 to Asn246), suggesting that the end of the C-terminal region is still highly mobile due to the lack of interactions between it and the acceptor substrate. In other words, it does not participate in acceptor binding.

When stored in the absence of a reducing agent, BoGT6a was observed as a disulphide-linked dimer (approximately 60 kDa) on SDS-PAGE. This resulted in protein aggregation and a reduction in enzyme activity (Tumbale and Brew, 2009). However, as discussed above, no disulphide bonds were observed in the structure of BoGT6a in complex with FAL. There were hydrophobic interactions between the 4 chains of the complex structure, namely residue Leu219 of chain B interacts with residue Asn127 of chain A (the same as in chain D and chain C), residue Glu223 and Pro221 of chain A interact with residue Glu216 and Gly217 of chain D respectively (the same as in chain C and chain B). In addition, there are hydrophobic interactions occurring symmetrically between chain A and chain C, and chain B and chain D (Figure 27B). Residues Glu132 and Lys128 of chain B also interact with the ligand of chain A (also in chain D and chain C) (Figure 29). However the protein is fully active as a monomer (Tumbale and Brew, 2009), and the distance between Lys128 of chain B and the O5' of FAL is too small to have a biological role. PISA analysis also indicated that the close contact between chain A and chain B, which solvent accessible area interface is 3.0% and 3.2% respectively, is not a biological contact, resulting instead from crystal packing. The same result was obtained for chain C and chain D which their solvent accessible area interface were 3.0% and 2.8% respectively.

Comparing the acceptor binding sites of the GT6 members, represented by GTA (PDB 1ZI3), GTB (PDB 1ZI1), bovine α3GT (PDB 1GX4) and BoGT6a in complex with FAL, shows a high conservation in both composition and conformation. The acceptor binding sites contain many similar residues such as Trp189, Thr134, Tyr153, His122, and Glu192 (BoGT6a numbering) (Figure 31), suggesting that the conformation of the acceptor binding site is conserved in the GT6 family, regardless of metal-independence or dependence. This also suggests that the metal ion is mainly involved in donor, rather than acceptor, binding.

In addition, comparing the acceptor pockets of these GT6 members illustrates structural conservation of many hydrophobic residues. For example, Trp189 corresponds to Trp314 of bovine α3GT and Trp300 of GTA and GTB. Pro123 and Phe125 are conserved in GTA and GTB (Pro234 and Phe236 respectively), but are replaced by larger hydrophobic residues (Ala248 and Trp250 respectively) in α3GT (Figure 31).

Previous structural studies have shown that Trp314 of bovine α3GT forms mainly hydrophobic interactions with the β-galactosyl moiety of the acceptor. The N of its indole ring is also within H-bonding distance of O6 of the glucose moiety of the acceptor substrate, lactose (Boix *et al.*, 2002, Zhang *et al.*, 2004). The mutant Trp314Tyr, lacking the N of the indole ring, displays a 30 fold decrease in its glycosyltransferase activity. This is in spite of the fact that both the acceptor and donor substrates retain the same orientations in the active sites and that affinity for the acceptor is 2 fold greater (Boix *et al.*, 2002, Zhang *et al.*, 2004). The mutation did not however affect the hydrolysis activity (Zhang *et al.*, 2004). This means that this residue plays a more important role in acceptor binding than in donor binding. These findings suggested that this residue plays a role in stabilising the bisubstrate transition state for the transfer of galactose to the carbohydrate, rather than to the water molecule. Such an important role explains why this Trp residue is well conserved among GT6 members.

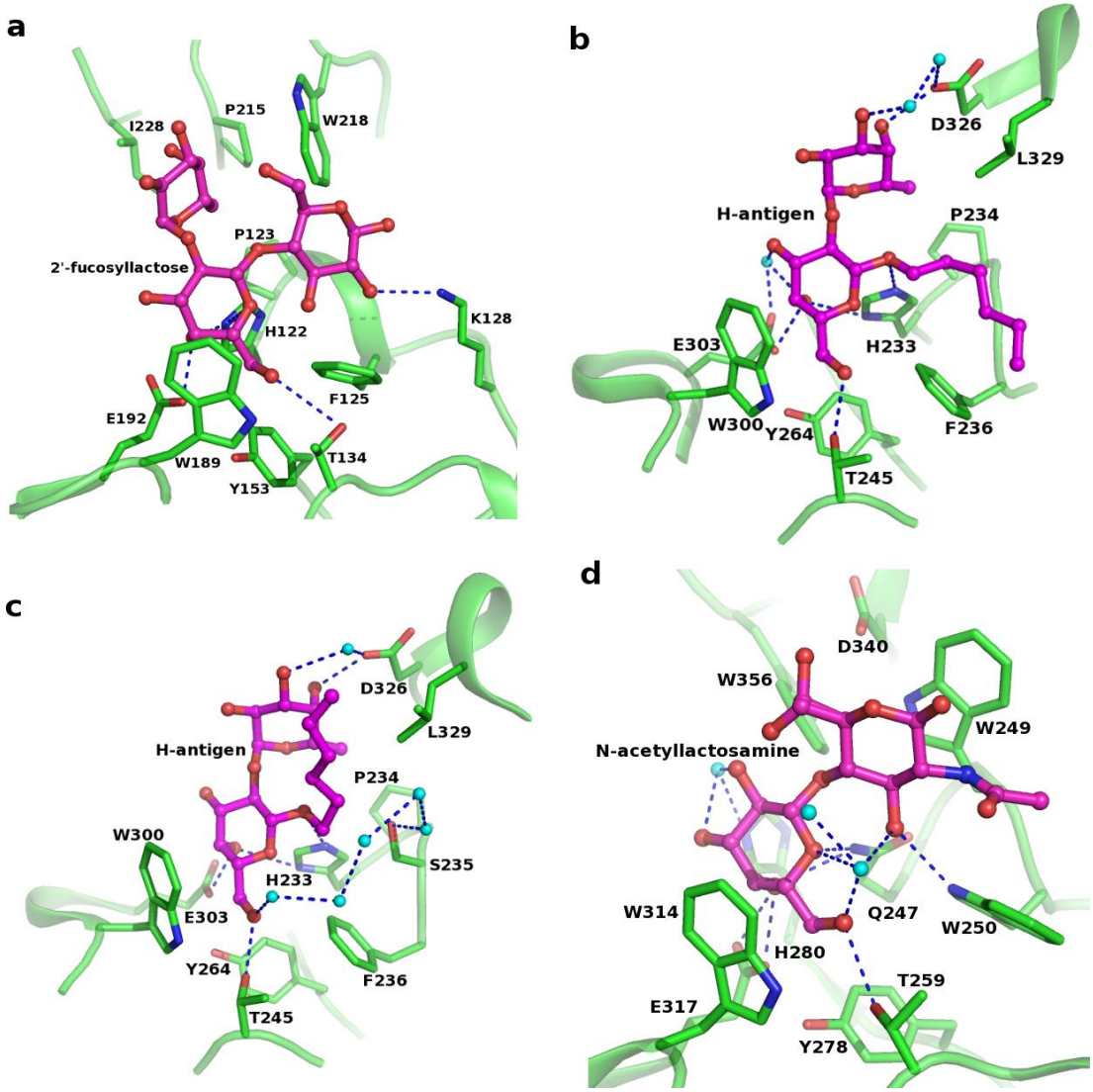


Figure 31. Acceptor binding pocket among GT6 family members showing the conserved residues and interactions. (a) BoGT6a in complex with FAL; (b) GTA in complex with H-antigen (PDB id: 1LZI); (c) GTB in complex with H-antigen (PDB id: 1LZJ); (d) α3GT in complex with LacNAc (PDB id:1GX4). Water molecules are shown as cyan spheres. Picture created using Pymol.

The bovine α3GT mutant Trp250Phe displays slightly increased transferase as well as hydrolysis activity. However, the affinity of the mutant for UDP-Gal is significantly reduced (approximately 10 fold) and the affinity for the acceptor, lactose, is also slightly decreased in the transferase reaction. Conversely, there was no noticeable change in substrate affinity in the hydrolysis reaction (Zhang *et al.*, 2004). The activity of the bovine α3GT mutant Trp250Tyr, was also affected in the same way; with a smaller increase in glycosyltransferase activity and a larger

increase in hydrolysis activity. In addition, both the mutants of bovine α3GT did not exhibit any selectivity between Lac and LacNAc (Zhang *et al.*, 2004). This suggests that the residue Trp250 only has a role in providing hydrophobic interactions with the carbohydrate substrate, helping the enzyme to accommodate it, but not in the catalytic activity of the enzyme. This also explained the differences observed at this position among GT6 members. The substitution of Trp250 in bovine α3GT by Phe125 in BoGT6a and Phe236 in GTA/GTB does not affect the activities of the enzymes as long as they are able to maintain hydrophobic interactions with the acceptor substrates.

Bovine α3GT was not able to accommodate an acceptor with a fucosyl derived substrate as GTA/GTB and BoGT6a were. In the acceptor binding site, a bulky residue, Trp356, forms hydrophobic interactions with the Gal moiety of its acceptor. BoGT6a and GTA/GTB have Ile228 and Ala343 respectively at this position (not shown in the figure as the small size makes it very difficult to see). These are smaller residues and hence are able to accommodate an additional sugar moiety or larger acceptor substrate (Figure 31).

The Trp356Thr mutation in bovine α3GT reduced the glycosyltransferase activity by 10 fold. This corresponds with an approximately 10-fold reduction in acceptor affinity. In contrast, although the affinity for the UDP-Gal substrate was reduced 2 fold, the hydrolysis capability of the mutant was reduced 16 fold. This finding also suggests that hydrophobic residues play an important role in the acceptor binding site.

In summary, in spite of a major functional divergence from vertebrate GT6s in having metal-independent catalytic activity, BoGT6a is strikingly similar to its mammalian homologues and also utilises precisely equivalent residues for binding acceptor substrates. This suggests the metal ion only affects the donor substrate binding of GT6 members through the interactions with the DXD motif. The conservation of a high density of hydrophobic residues allows these enzymes to accommodate carbohydrate substrates, but the difference in the sizes of these residues may be involved in the specification of the different sizes of their acceptors.

CHAPTER III

Structure of BoGT6a E192Q in complex with UDP-GalNAc

3 Structure of BoGT6a E192Q in complex with its donor substrate UDP-GalNAc

3.1 Introduction

One of the main aims of this project was to shed light into the structural basis of the BoGT6a catalytic mechanism. Previous structural studies on GT6 members indicate that Glu317 is a potential catalytic nucleophile of bovine α3GT, and Glu303 is a potential catalytic nucleophile of GTA and GTB (Gómez *et al.*, 2013, Gómez *et al.*, 2012, Soya *et al.*, 2011, Monegal and Planas, 2006). Both of these belong to LBR-F, and may be involved in a double displacement mechanism. These residues are highly conserved in the GT6 family, and their derived mutants show a significant reduction in their activity (Zhang *et al.*, 2003, Patenaude *et al.*, 2002).

Superposition of the structure of BoGT6a with those of its homologues shows that Glu192, which corresponds to Glu317 of bovine α3GT, can be a catalytic nucleophile in BoGT6a catalytic activity. Kinetic assay of BoGT6a E192Q has shown that glycosyl transferase activity of this mutant is reduced by a factor of 30000 compared to that of wild type enzyme, implicating its importance in enzyme activity (Tumbale and Brew, 2009).

To understand more about the role of Glu192 in BoGT6a catalytic activity, crystallisation attempts of the E192Q mutant, in complex with substrates, were performed. This chapter describes the crystallisation, structure determination and analysis of this mutant with its active donor substrate UDP-GalNAc bound in the active site. In addition, the donor bound complexes of the mutant explain how the residues of BoGT6a function in the absence of metal ions.

3.2 Methods

3.2.1 Protein preparation

The protein BoGT6a E192Q, was received from Professor Keith Brew, Florida Atlantic University, USA, in a number of different buffers at concentrations ranging from 0.3-0.4 mg/ml (Table 3). The protein from each batch needed to be

concentrated and this was achieved by centrifugation at 4 ^{o}C, 4000 rpm for cycles of 45 minutes using a Thermo Scientific Heracus Megafuge 16R Centrifuge, until the concentration reached about 8 mg/ml. An Amicon Ultra-15 MW3000 (Millipore) was utilised during this process. Protein concentration was calculated using a Nanodrop 2000c spectrophotometer (Thermo Scientific) to measure the absorbance at 280 nm with an absorbance coefficient of 1.4 for 1 mg/ml, calculated using ExPASy – ProtParam (Expert Protein Analysis System) tool (Gasteiger *et al.*, 2003).

Table 3. Buffer conditions and concentrations of the BoGT6a E192Q batches supplied

Batch	Buffer	Concentration (mg/ml)
1a	20mM Tris-HCl, 0.5M NaCl, 1mM DTT, pH 7.5	0.33
1b	20mM Tris-HCl, 0.5M NaCl, 1mM DTT, pH 7.9	0.40
2	20mM Tris-HCl, 0.5M NaCl, 150mM DTT, pH 7.9	0.36
3a	20mM Tris-HCl, 0.5M NaCl, 150mM Imidazole, pH 7.9	0.30
3b	20mM Tris-HCl, 0.1M NaCl, 2mM DTT, 10mM EDTA, pH 7.0	0.30
4	20mM Tris-HCl, 0.1M NaCl, 2mM DTT, 10mM EDTA, pH 7.0	0.27

To make the complex of BoGT6a E192Q with the donor substrate UDP-GalNAc, the concentrated protein was incubated with UDP-GalNAc at a final concentration of 10 mM at room temperature.

For the complex of BoGT6a E192Q with both the donor UDP-GalNAc and the acceptor 2'-fucosyllactose (FAL), the concentrated protein was mixed with UDP-GalNAc and FAL simultaneously, both at a final concentration of 10 mM at room temperature.

All the complexes were incubated at 4 ^{o}C overnight prior to the crystallisation trials.

3.2.2 Crystallisation

A Phoenix crystallisation robot (Art Robbins Instruments) was used to set up commercial crystallisation screens using 96-well Intelli-plates ®. Screens used included: Structure screen I & II, Clear Strategy Screen I, Clear Strategy Screen II, PACT premier, and Proplex (Molecular Dimensions). The sitting drop vapour diffusion method was used with a drop size of 0.2 µl and a 1:1 ratio of protein: reservoir solution. The plates were set up at room temperature, and then incubated at 16 ^{o}C.

Subsequently, crystallisation was scaled up to 24-well plates to optimise the initial "hits" from the 96-well screens. During crystallisation optimisation, incubation time for the complex of BoGT6a E192Q with UDP-GalNAc was reduced to 1 hour at room temperature prior to crystallisation. Precipitant concentrations, as well as the type of PEG and buffer used were varied, but parameters such as pH, incubation temperature, protein concentration and the ratio of protein to reservoir solution were kept constant. Crystals obtained from both screening and optimisation were analysed at Diamond Light Source (Didcot, Oxon-UK).

3.2.3 Structure determination

Diffraction datasets for the complexes were collected at Diamond Light Source (Didcot, Oxon, UK) at 100 K. Cryo cooling was achieved by stabilising the crystals (prior to X-ray data collection) in 25%v/v Glycerol. Datasets were processed automatically by XIA2 (Winter, 2010) available at Diamond Light Source.

Three datasets were used to determine structures of BoGT6a E192Q in complex with its donor substrate UDP-GalNAc. The first was dataset 2, which was processed at 3.50 Å in space group $P2_1$. The second was dataset 4, which was processed at 3.42 Å in space group $P2_12_12_1$. The last was dataset 5, which was processed at 2.78 Å in space group $P2_12_12_1$.

Phases for the BoGT6a E192Q-UDP-GalNAc complex at 2.78 Å were calculated by the MR method (using the native BoGT6a structure as the starting model) with four molecules per asymmetric unit using Phaser (program version 2.5.0) in PHENIX

software suite (version 1.8.0). The missing loop (residues 126 – 150) in the BoGT6a structure was built based on the observed electron density map. Electron density was only observed for α-GalNAc (noted as GalNAc), and so only GalNAc, rather than UDP-GalNAc, was inserted into the structure. Further refinement and model building were performed using the PHENIX software suite and COOT. This structure is henceforth referred to as BoGT6a E192Q-GalNAc or form I structure.

Chain A of the BoGT6a E192Q•GalNAc structure without the GalNAc moiety was used as a starting model for phase calculation of the other datasets. As in the BoGT6a E192Q.GalNAc structure, there are 4 molecules per asymmetric unit in the protein structure from dataset 4, whilst there are 16 molecules per asymmetric unit in the protein structure from dataset 2. UDP-GalNAc, UDP and GalNAc were inserted into the protein structures based on their observed electron densities. Final structures were achieved after several refinement cycles using PHENIX software suite and COOT. The structure obtained from dataset 4 was named BoGT6a E192Q•UDP-GalNAc structure in orthorhombic form or form II structure, and the structure from dataset 2 is referred to as BoGT6a E192Q•UDP-GalNAc structure in monoclinic form or form III structure.

3.3 Results

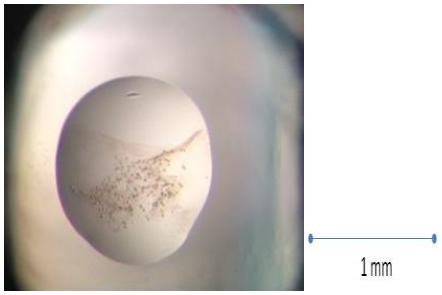
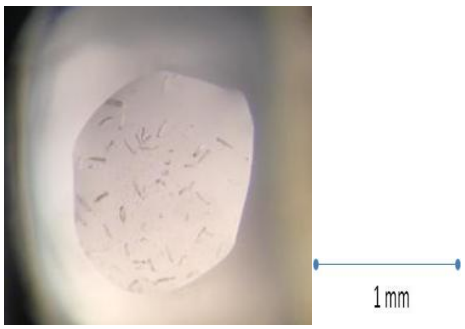
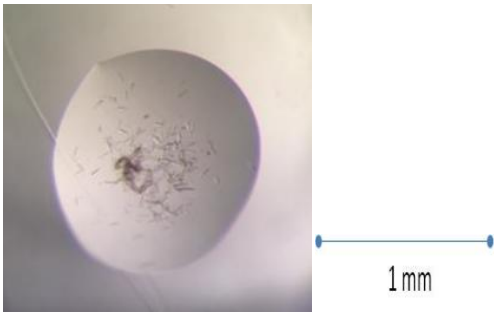
3.3.1 Crystallisation

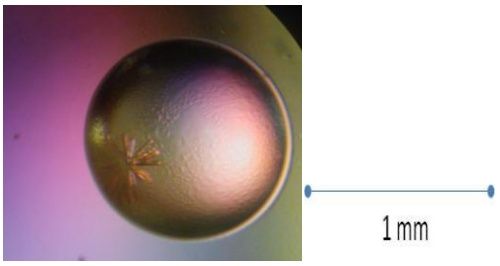
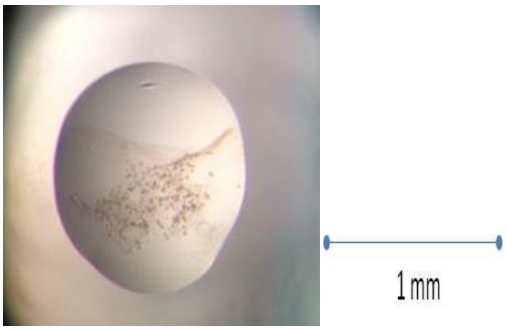
The BoGT6a E192Q protein was not stable in buffer at pH 7.5 and pH 7.9 which caused precipitation during transportation. This led to a limited protein source for crystallisation screens. The crystallisation screens for the complexes derived from these proteins did not provide any hits, only precipitation. The BoGT6a E192Q protein from batch 3b was more stable and some “hits” were obtained from its crystallisation screens (Table 4).

Although BoGT6a E192Q was co-crystallised with 10mM of UDP-GalNAc and with 10 mM of both UDP-GalNAc and FAL, only crystals of BoGT6a E192Q in complex with UDP-GalNAc were obtained from the commercial screens. The crystallisation trials for the ternary complex only showed aggregation and precipitation. Lower protein concentration (6 mg/ml) was tried but the result was not improved. As the

protein source was limited, the priority was to use BoGT6a E192Q to set up crystallisations for the donor bound complex, as this had yielded some "hits" during the condition screening experiments.

Table 4. Crystallisation screen results for BoGT6a E192Q batch 3b in complex with UDP-GalNAc

Name of commercial screen and Condition	Crystal form
Crystal Strategy Screen I MD1-31 (D3) 0.2M $MgCl_2$ 0.1M sodium cacodylate, pH6.5 10% PEG 8000 + 10% PEG 1000	
Crystal Strategy Screen I MD1-31 (C2) 0.2M Li_2SO_4 0.1M sodium cacodylate, pH6.5 25% PEG 2000 MME	
Structure Screen 1 & 2 HT-96 MD1-30 (G2) 0.2M $(NH4)_2SO_4$ 0.1M MES, pH 6.5 30% PEG 5000 MME	

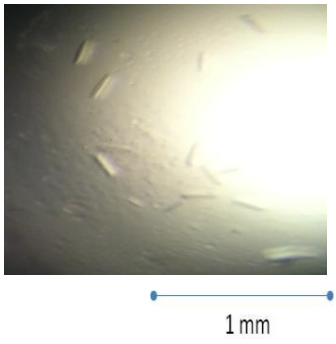
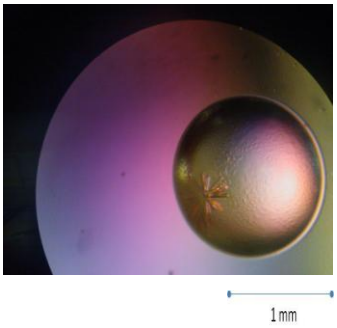
Name of commercial screen and Condition	Crystal form
ProPlex screen HT-96 MD1-42 (E12) 0.2M $(NH_4)_2SO_4$ 0.1M MES, pH 6.5 20% PEG 8000	
Crystal Strategy Screen II HT-96 MD1-32 (E6) 0.2M $Ca(CH_3COO)_2$ 0.1M Tris-HCl, pH 7.5 15% PEG 4000	

Most of the hits were of insufficient quality for mounting, except for one cluster of bar-shaped crystals. These appeared after 1 month of incubation at 16 ^{o}C in a well solution containing: 20% PEG 3350, 0.1M Na citrate, pH 5.0, 20% PEG 8000 from the Proplex crystallisation screen (Molecular Dimensions Ltd., UK). These crystals diffracted to 2.78 Å (Dataset 5, Table 5). Unfortunately, attempts at repetition and optimisation of this condition did not yield any good crystals.

Based on the hits obtained (Table 4), modified crystallisations were set up manually in 24-well plates using the vapour diffusion, hanging drop method. Analysis of the hit conditions indicated that the complex of BoGT6a E192Q and UDP-GalNAc tended to form crystals with Li_2SO_4 or $(NH_4)_2SO_4$ at pH 5.6 and pH 6.5. Different PEGs, including PEG 4000, PEG 3500, PEG 6000 and PEG 8000, and different buffers at pH 5.5 and pH 6.5 were tried. 0.2 M of either Li_2SO_4 or $(NH_4)_2SO_4$ in 0.1 M Bis Tris, pH 5.5 and 20 % PEG 3350 provided potential conditions for the complex, which produced many bar-shaped crystals diffracting to between 3.30 Å to 4.03 Å. Two representatives of the crystals from these conditions, which gave

sufficient quality diffraction data for protein structure determination are shown in Table 5 (Dataset 2 and Dataset 4).

Table 5. Crystallisation conditions of the diffracted crystals of BoGT6a E192Q in complex with UDP-GalNAc

Conditions	Pictures	Results
0.2M Li_2SO_4 0.1M Bis Tris, pH 5.5 20% PEG 3350		Dataset 2
0.2M $(NH_4)_2SO_4$ 0.1M Bis Tris, pH 5.5 20% PEG 3350		Dataset 4
0.1M Na citrate, pH 5.0 20% PEG 8000		Dataset 5

3.3.2 Structure determination

The diffracting crystals from both screening and optimisation provided 8 datasets in total at different resolutions. Diffraction data, which were automatically indexed by XIA2 at Diamond Light Source, are summarised in Table 6. Only data collected at greater than 3.50 Å resolution; datasets 2, 4 and 5, were used because resolution lower than this is not useful for accurate depiction of substrate conformation.

Diffraction dataset 2 was collected at station I04, Diamond Light Source, using a ADSC Q315 detector with data collection parameters: $\lambda = 0.9795$ Å, $\Delta \varphi = 1.0^{o}$, and exposure = 2 seconds (Figure 32). 128 images were collected for this dataset. The highest resolution of these data was 3.35 Å, but it was processed at 3.50 Å for the best R_{merge} value (Table 6).

Diffraction datasets 4 and 5 were collected at station I04-1, Diamond Light Source, using a Pilatus 2M detector with data collection parameters: $\lambda = 0.9200$ Å, $\Delta \varphi = 1.0^{o}$, and exposure = 3 seconds (Figure 33 and Figure 34). 150 images were collected in total for each dataset. The datasets were processed at their highest resolutions: 3.42 Å for dataset 4 and 2.78 Å for dataset 5 (Table 6).

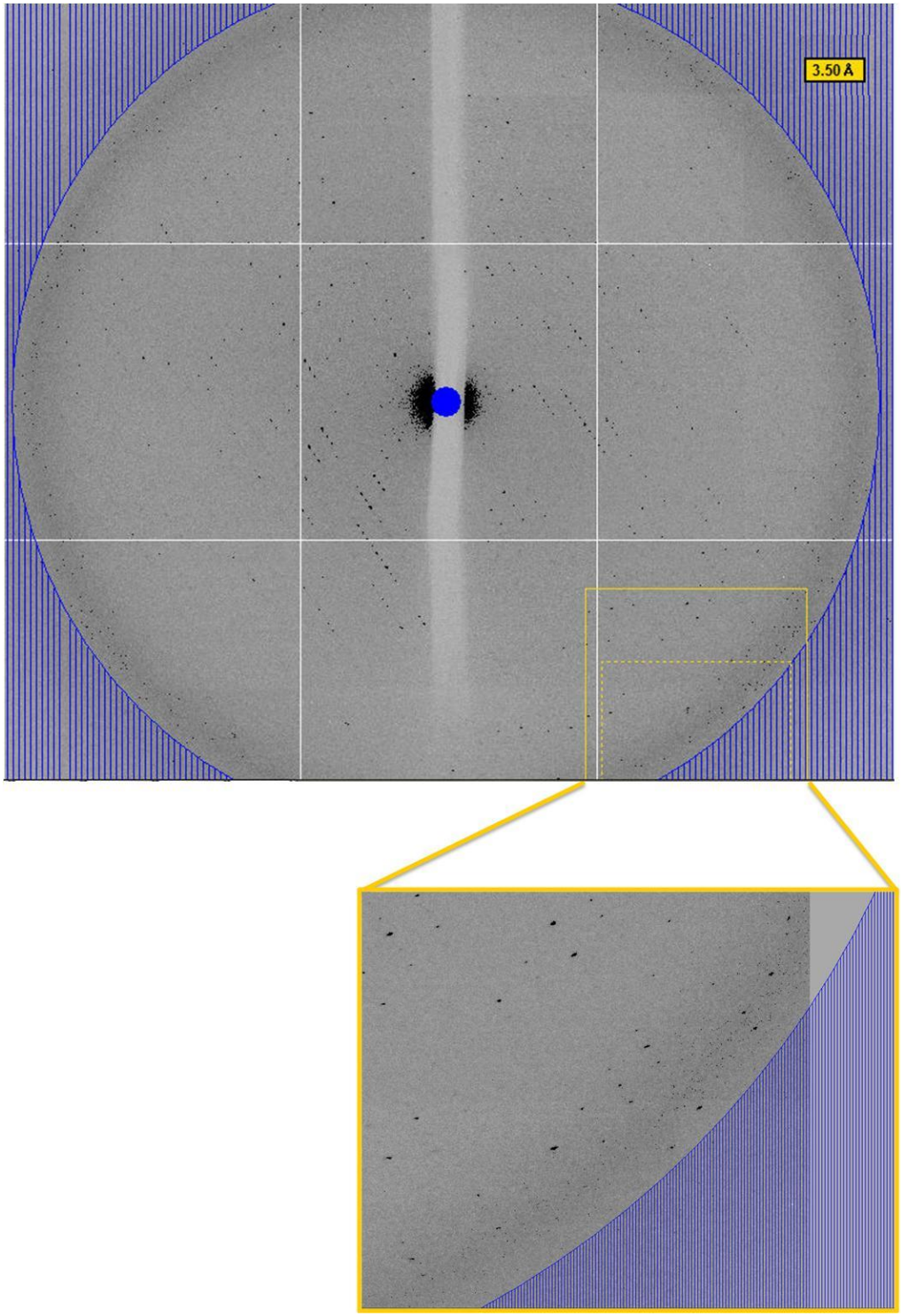


Figure 32. Diffraction image from the crystal of BoGT6a E192Q in complex with UDP-GalNAc that diffracted to 3.50 Å (dataset 2). The inset represents a portion of the image zoomed in to show the highest resolution spots.

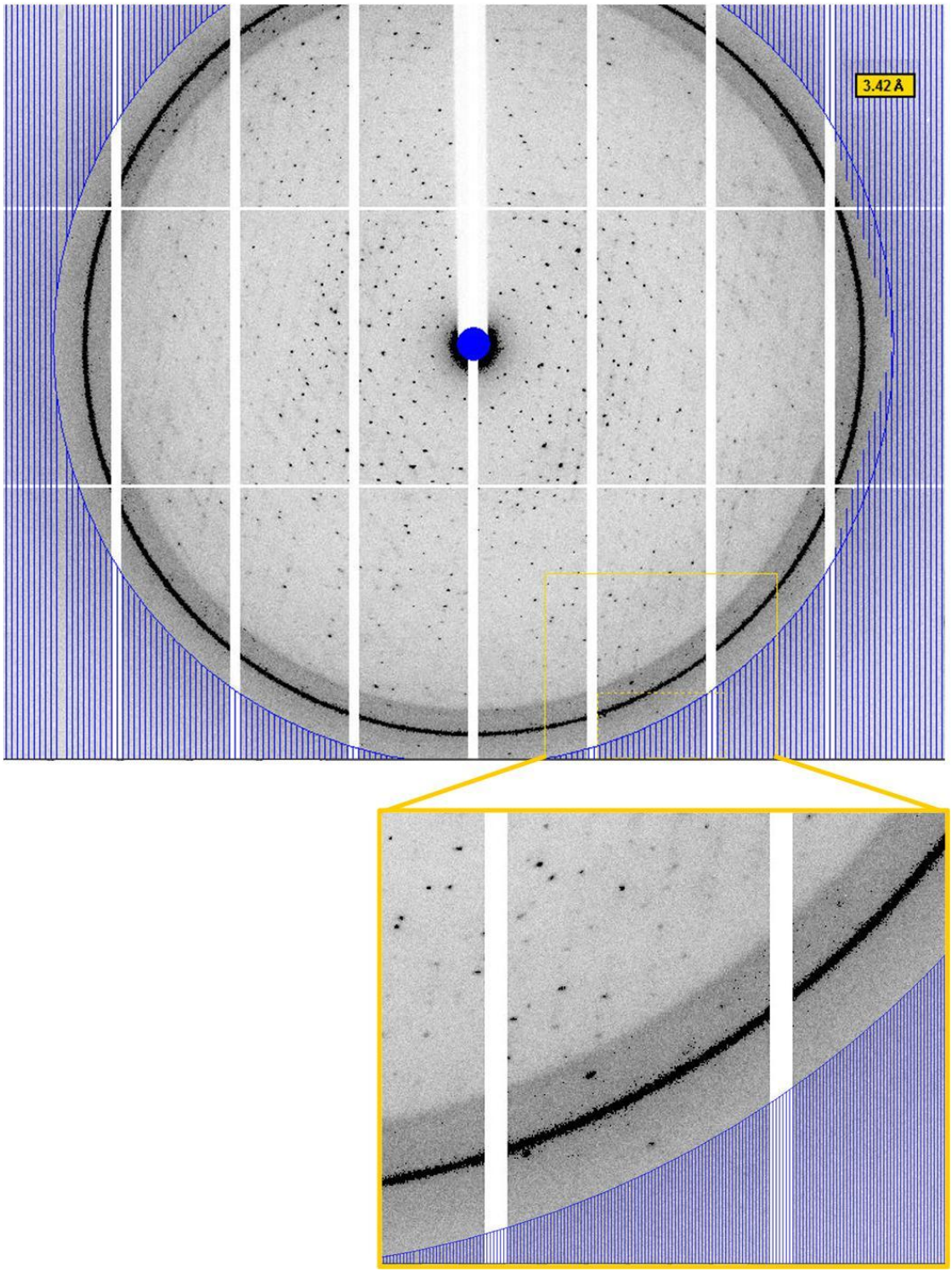


Figure 33. Diffraction image from the crystal of BoGT6a E192Q in complex with UDP-GalNAc that diffracted to 3.42 Å (dataset 4). The inset represents a portion of the image zoomed in to show the highest resolution spots.

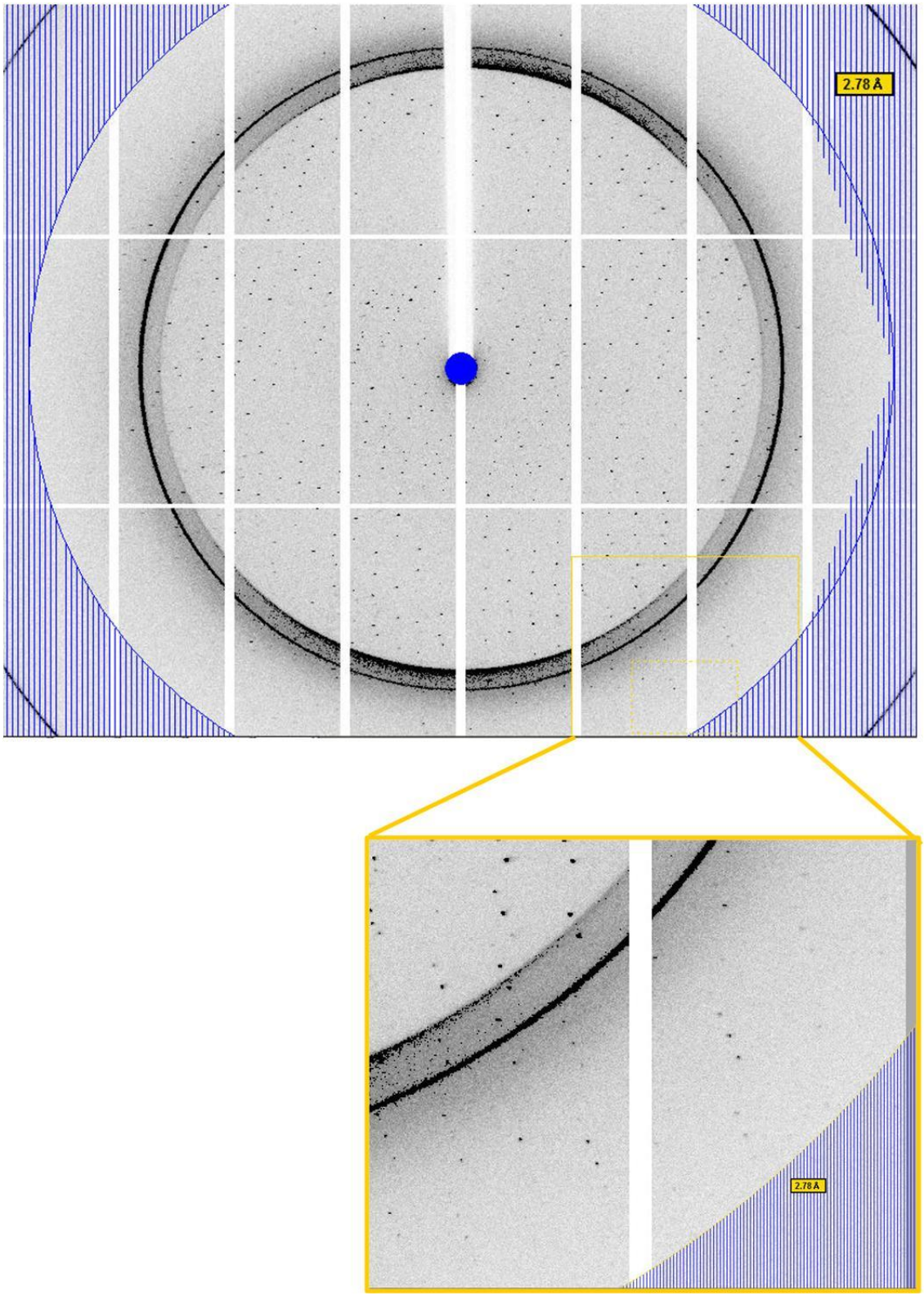


Figure 34. Diffraction image from the crystal of BoGT6a E192Q in complex with UDP-GalNAc that diffracted to 2.78 Å (dataset 5). The inset represents a portion of the image zoomed in to show the highest resolution spots.

Table 6. Information of data collections for BoGT6a E192Q in complex with UDP-GalNAc crystals

Dataset	Number of images	Resolution (Å)	Space group	Cell dimensions	R_{merge}	Completeness (%)
1_1	50	3.91	P2	a = 177.2 Å, b = 79.9 Å, c = 178.8 Å β = 94.4 ^{o}	0.067	69.20
1_2	100	3.91	P2	a = 177.0 Å, b = 79.8 Å, c = 178.6 Å β = 94.4 ^{o}	0.119	93.50
2	128	3.50	$P2_1$	a = 176.7 Å, b = 79.7 Å, c = 179.1 Å β = 95.2 ^{o}	0.134	98.00
3	150	3.75	$P2_12_12_1$	a = 80.1 Å, b = 119.8 Å, c = 131.7 Å α = β = γ = 90.0 ^{o}	0.087	99.70

Dataset	Number of images	Resolution (Å)	Space group	Cell dimensions	R_{merge}	Completeness (%)
4	150	3.42	$P2_12_12_1$	a = 80.1 Å, b = 120.2 Å, c = 131.8 Å α = β = γ = 90.0 ^{o}	0.085	97.60
5	150	2.78	$P2_12_12_1$	a = 80.1 Å, b = 115.6 Å, c = 126.1 Å α = β = γ = 90.0 ^{o}	0.095	97.70
6	120	3.81	$P2_12_12$	a = 120.1 Å, b = 130.2 Å, c = 79.3 Å α = β = γ = 90.0 ^{o}	0.122	94.30
7	250	4.03	$P2_1$	a = 176.1 Å, b = 79.8 Å, c = 178.3 Å β = 95.0 ^{o}	0.238	99.40
8	200	3.30	$P2_12_12$	a = 382.4 Å, b = 79.3 Å, c = 114.0 Å α = β = γ = 90.0 ^{o}	0.182	95.60

3.3.2.1 Structure of BoGT6a in complex with GalNAc (form I)

The processed results of dataset 5 were used to determine the structure of BoGT6a in complex with UDP-GalNAc. Matthews_Coef indicated that there were 4 molecules per asymmetric unit with a solvent content of 51.17 %. The MR method using Phaser version 2.5.3 in PHENIX was thus set as 4 components in an asymmetric unit with a 4-copy search. Chain A of the BoGT6a•FAL structure without the FAL moiety was used as a starting model. The solution gave a structure consisting of 4 molecules per asymmetric unit with LLG 4198 and TFZ 32.7.

Each molecule had difference strong electron density for a continuous structure from residue Met1 to Asp230. As expected, there was also difference electron density at the active site and the C-terminal region in each chain (Figure 35A, Figure 36A). The C-terminal regions were built based on their electron densities in each chain to get the final structure with 236 residues for chain A and chain B, 237 residues for chain C, and 235 residues for chain D (Figure 35B and C).

Although the difference density was observed in the active site of all 4 molecules, it was not sufficient for building the whole donor substrate, UDP-GalNAc, only individual GalNAc moieties. GalNAc moieties were added based on the positive electron density (Figure 36). The occupancies were different in the four chains: 0.65 in chain A, 0.70 in chain B, 0.77 in chain C, and 0.81 in chain D. This was due to their flexibility in the active sites. Since this structure was obtained from a crystal grown from the solution of BoGT6a E192Q in complex with UDP-GalNAc, the residue Glu192 in each chain was changed to Gln192, even though the electron density was not sufficient to distinguish between residue Gln and residue Glu (Figure 36). Water molecules were built into the structure during refinement cycles. Refinement was performed using the Refine program in PHENIX with NCS restraints and Coot until the values of R and R_{free} reached 23.14 and 27.35 respectively. 96.1 % of residues in this structure lie in the favoured region of the Ramachandran plot. Further crystallographic statistics are summarised in Table 7. The final structure, named BoGT6a E192Q•GalNAc structure (form I), has 4 molecules per asymmetric unit, each with GalNAc moieties in their active sites (Figure 37).

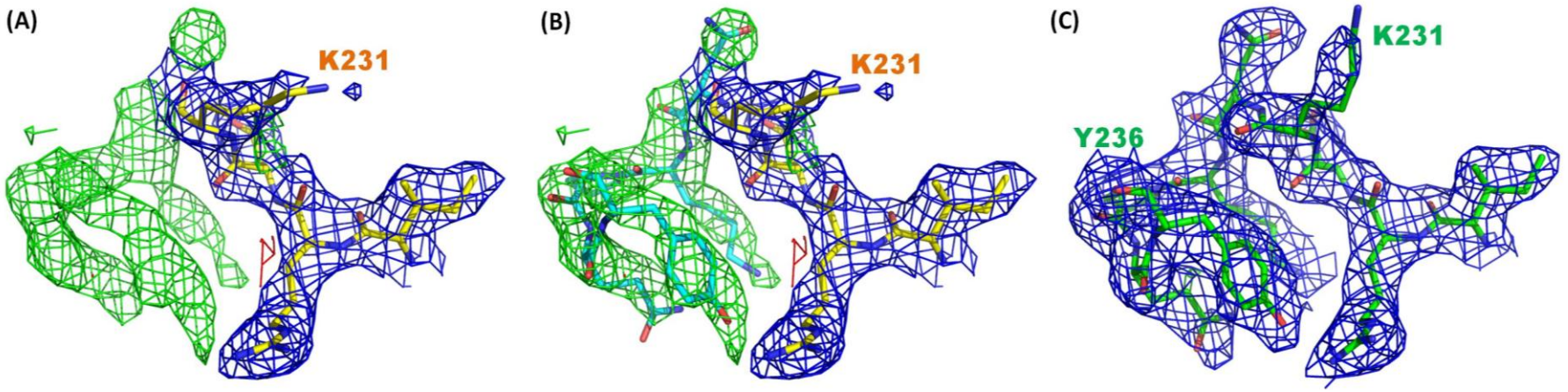


Figure 35. Electron densities of C-terminal region (of chain A as representative) of BoGT6a E192Q in complex with donor substrate derived from dataset 5. (A) the difference densities of the end residues of C-terminus (after Lys231) after phasing with the structure of chain A of BoGT6a•FAL as a starting model. (B) shows the end residues of the C terminus built based on the positive difference electron density map. (C) shows the final electron densities fitted well with the end of the C-terminal region. Protein was shown as line which was coloured in yellow for the structure after the first search and in green for the final structure. The $2F_o$-F_c map is contoured at 1σ and coloured in blue. The F_o-F_c is contoured at 3 σ and coloured in green for positive electron density and in red for negative electron density. Residues are noted in 1 letter abbreviation. The picture was created by using Pymol.

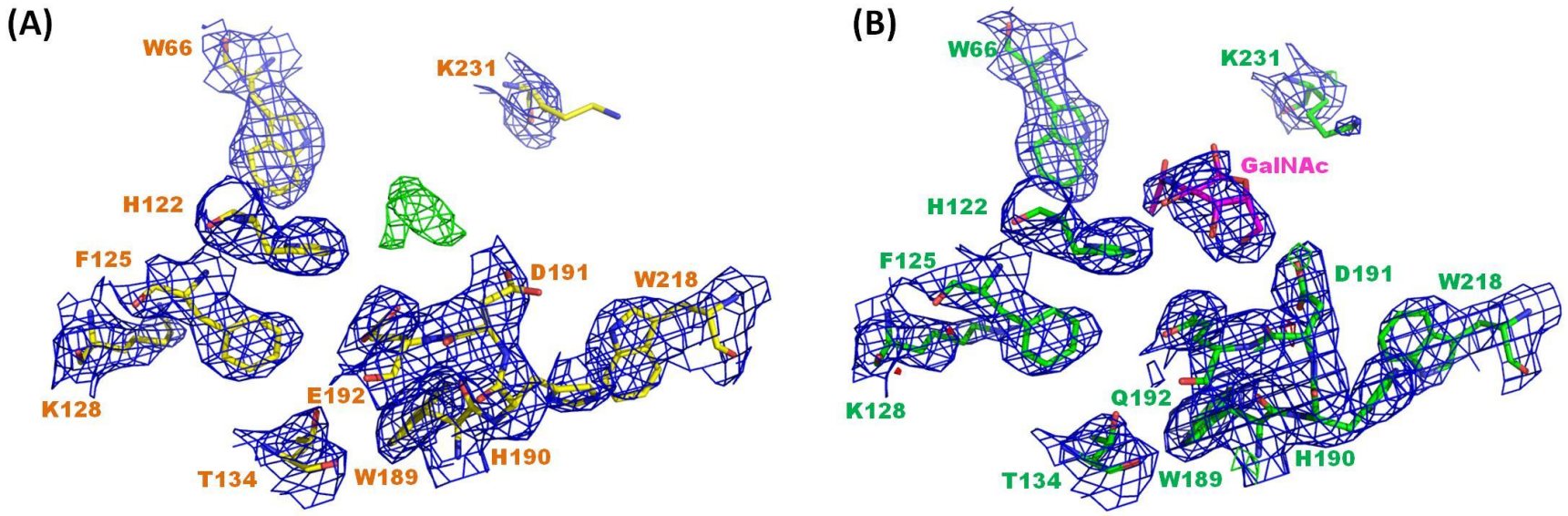


Figure 36. Electron densities of ligand (of chain A as representative) in the structure BoGT6a E192Q in complex with donor substrate derived from dataset 5. (A) the different densities appeared in the active site of BoGT6a E192Q after phasing with the structure of chain A of BoGT6a•FAL as a starting model. (B) shows the final electron densities fitted well with GalNAc moiety. Protein is shown as line which is coloured in yellow for the structure after the first search and in green for the final structure. The ligand is shown as line in magenta. The $2F_o$-F_c map is contoured at 1σ and coloured in blue. The F_o-F_c is contoured at 3 σ and coloured in green for positive electron density and in red for negative electron density. Residues are noted in 1 letter abbreviation, showing the active site of the enzyme. The picture was created by using Pymol.

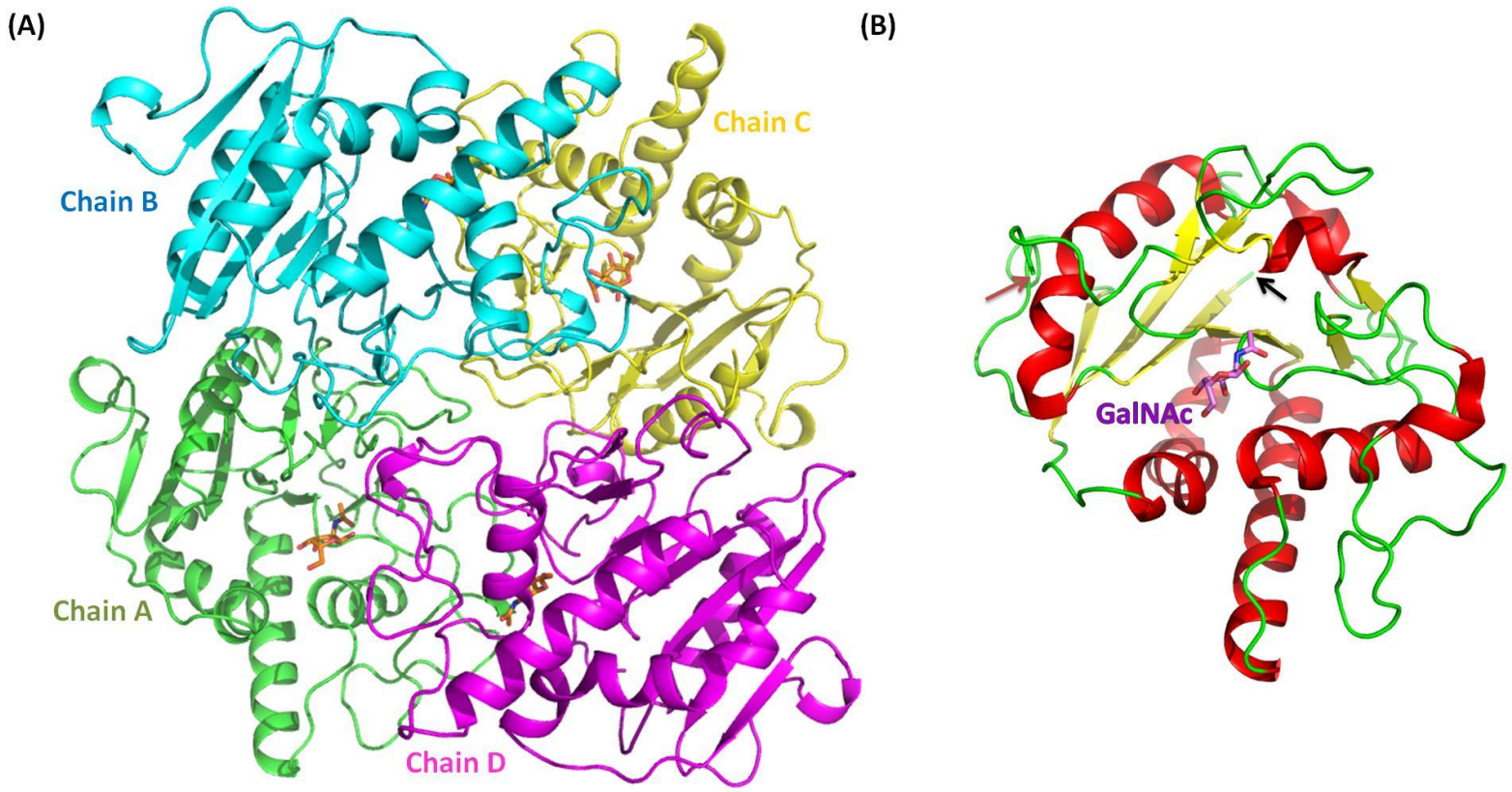


Figure 37. Crystal structure of BoGT6a in complex with β-GalNAc from the orthorhombic crystal form (form I). (A) 4 chains in a asymmetric unit. Protein is shown in cartoon representation and coloured by chain. The ligands are shown as orange sticks. (B) one chain as representative. Protein is shown in cartoon representation and coloured by secondary structure. The ligand is shown as stick in violet and labelled as GalNAc. Red arrow indicates the C-term, and black arrow the N-term. Picture created using Pymol.

3.3.2.2 Structure of BoGT6a in complex with UDP-GalNAc (form III)

The first datasets obtained were actually dataset 1 and dataset 2, but only dataset 2 was at a resolution sufficient for protein structure determination, 3.50 Å. Automatic indexing by XIA2 at Diamond Light Source showed that the crystal belonged to spacegroup $P2_1$ with R_{merge} 0.134 and 98 % completeness. The Matthew's coefficient and solvent content calculated from the unit cell dimensions and the molecular weight of BoGT6a without His-tag weight (Mw 29000) using MATTHEWS_COEF indicated that the unit cell contained 16 molecules per asymmetric unit with a solvent content of 54.66 %. Although the unusually high number of molecules in the asymmetric unit suggested that the symmetry of the crystal could be higher than $P2_1$, Pointless gave the probability of $P2_1$ as 95 %. Furthermore, attempts to manually index dataset 2 in $P2_12_12_1$, C222 or $C222_1$ using Imosflm failed at the scaling step.

The processed data from XIA2 was used for further analysis. The MR method using Phaser (program version 2.3.0) in CCP4i (version 6.2.0) was applied to solve the phase problem with BoGT6a in complex with FAL using chain A of the structure without FAL as a starting model. Although the first search was set with 16 molecules per asymmetric and 16 copies search, the search failed to find a complete solution. A partial solution with only 15 molecules in the asymmetric unit was found with LLG and TFZ values of 12443 and 48.2 respectively. A second search was performed with all parameters was set as in the first search but only 1 copy search. This gave 1 molecule with LLG 158 and TFZ 9.3. Similar searching steps were repeated in which each step was set with 1 copy search each time utilising the solution from the previous search. After each search, one more molecule in the asymmetric unit was found and the LLG value was increased. However, the last search for 16^{th} molecule failed as the first search had, even though there was difference electron density visible for one more molecule in the asymmetric unit (Figure 38).

MR was performed again with Phaser program in PHENIX using the same processed data from XIA2, again setting 16 chains in the asymmetric unit and 16 copy search with chain A of the structure of BoGT6a•FAL without FAL moiety as a starting model. Unlike the CCP4i result, the solution from PHENIX was completed with 16 molecules found and there was no large difference electron density in the map,

except for the ligands in the active site and the C-terminal region of each chain. The values of LLG and TFZ were 12268 and 24.4 respectively. Each chain of the model also had continuous structure from Met1 to Lys231, and the residue Glu192 was kept as in the starting model. The first refinement gave R and R_{free} values of 25.43 and 30.22 respectively.

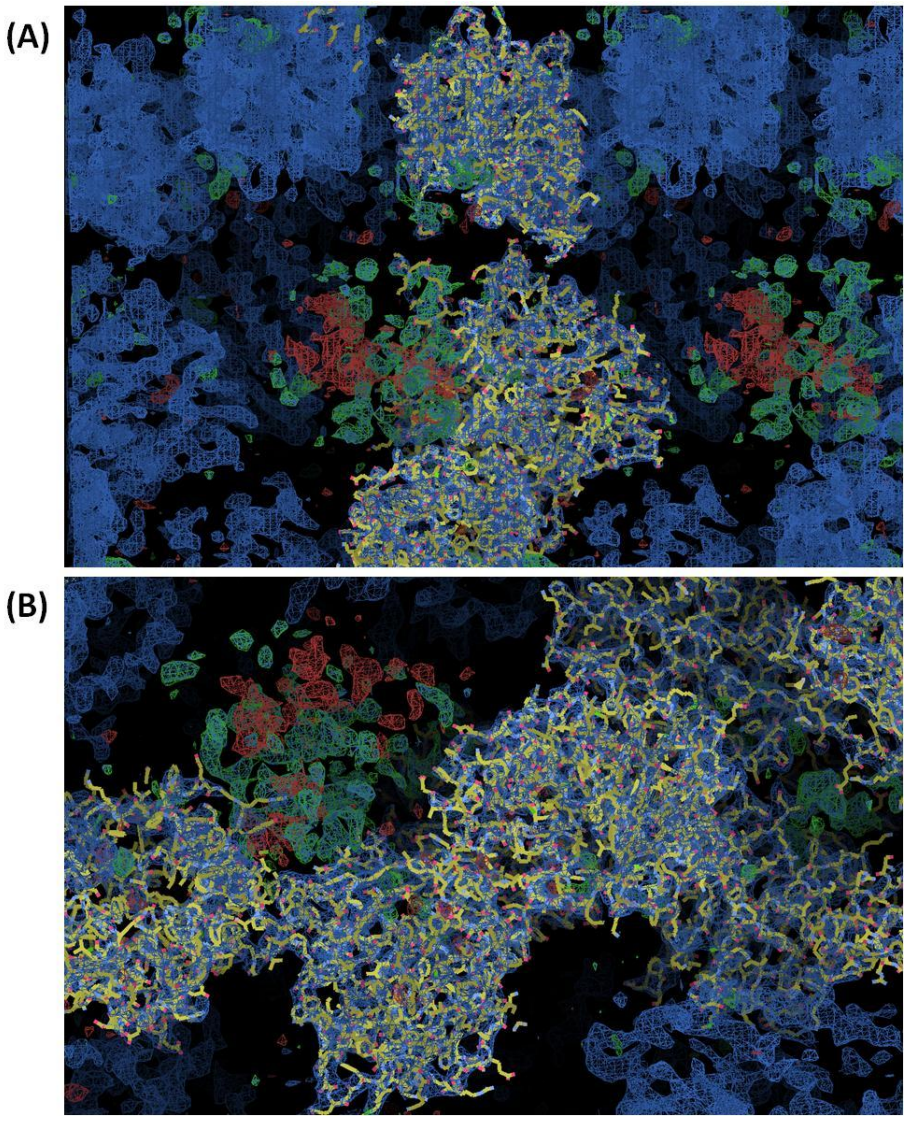


Figure 38. Electron densities for missing molecules in asymmetric unit after MR searching using Phaser_MR program in CCP4i version 6.2.0. (A) two difference electron densities appeared in the map after 14 searches. (B) one difference electron density presents in the map after 15 searches. The maps are set at 80 Å diameter. The $2F_o$-F_c map is contoured at 1σ and coloured in blue. The F_o-F_c is contoured at 3 σ and coloured in green for positive electron density and in red for negative electron density. The picture was created by using Coot.

By the time the solution was found using PHENIX, datasets 4 and 5 had been collected. After solving the structure of BoGT6a E192Q•GalNAc from dataset 5, chain A from this structure, without the GAL moiety, was used as the starting model for phasing of dataset 2. The structure must be a better starting model than BoGT6a•FAL structure because it derived from the same protein and donor substrate. Its conformation was expected to be more similar to the target structure. The full solution also found 16 molecules per asymmetric unit but the values of LLG and TFZ were slightly higher than the previous search with BoGT6a•FAL as searching model; 13278 and 25.7 respectively. Although the general topology of the structures and the crystal packing from both searches were similar, the structure found by the MR method with BoGT6a•GalNAc as a starting model was used for further analysis. This was because not only were the LLG and TFZ values better, but each chain also had a longer C-terminal region, ending at Tyr236, as in the model. Since the mutant structure was used as the search model, each chain of the resulting structure also contained the E192Q mutation. The first refinement improved the R and R_{free} values, which were 25.78 and 29.77 respectively.

Like the BoGT6a•GalNAc structure, there was also positive difference electron density present in the active sites and the C-terminal regions of all chains, but these densities were not consistent in all chains. At the C-terminal region, chains A, B, C, D, I, J, K, and L did not have sufficient positive electron density to build further, whilst the other chains (E, F, G, H, M, N, O and P) had larger densities (Figure 39). Residues of the C-terminus were added to each chain based on their electron densities. After a few refinement cycles, chains A, B, C, D, I, J, K and L ended with Tyr236. Chains E, F, G, H and M ended at Lys245, chains N and P at 243, and chain O at Glu240. In addition, a small amount of positive electron density was visible at the N-termini of chains A, B, C, D, E, F, G, H, and K. In accordance with the sequence of BoGT6a, residues of the His-tag were added and the final structure contained residue His0 and Ser-1 in chains A, B, C, H, and only His0 in chains D, E, F, G and K (Figure 40).

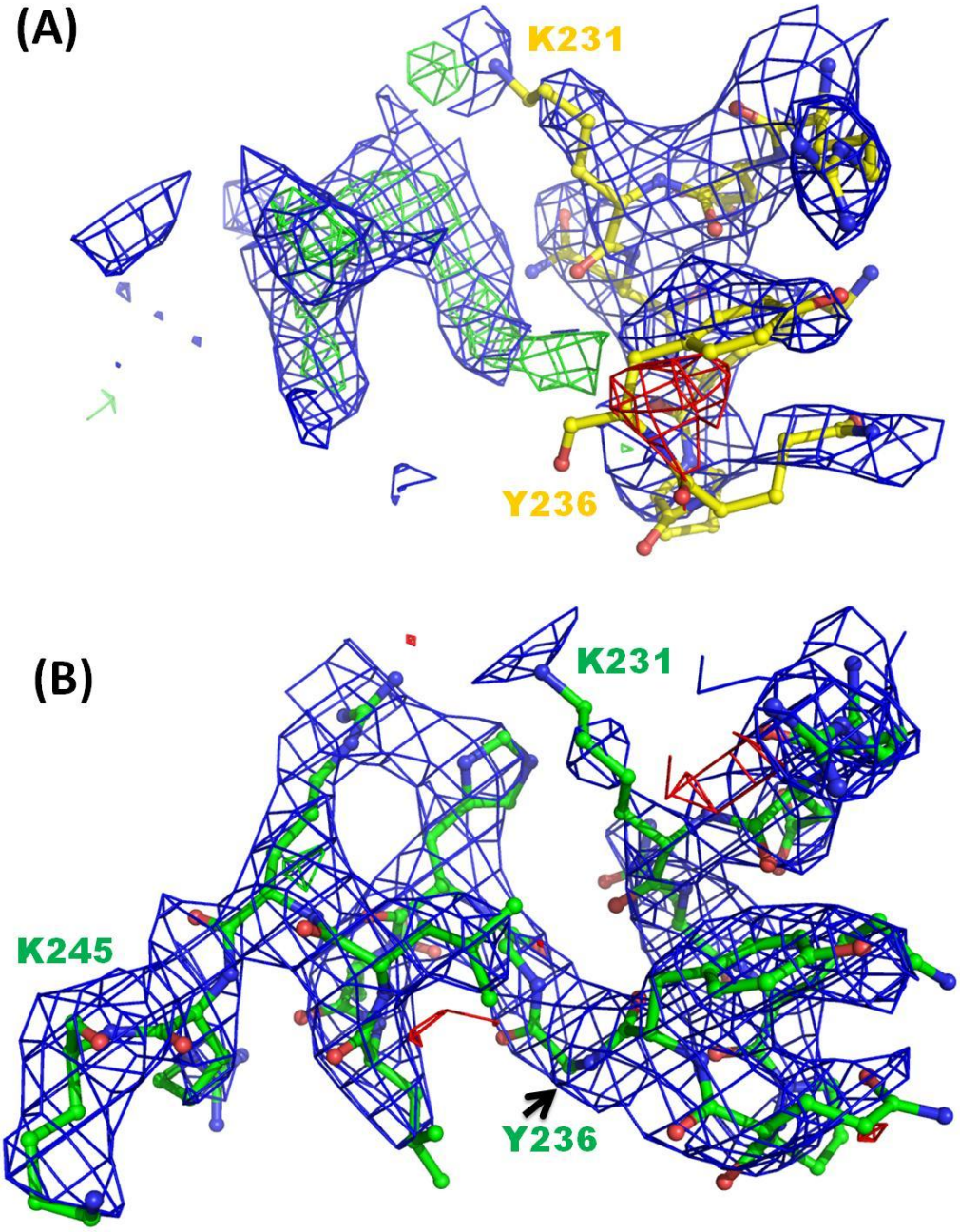


Figure 39. The electron density maps of the C-terminal region (of chain E as representative) of the structure BoGT6a in complex with UDP-GalNAc (derived from dataset 2) before and after the missing residues were built. (A) the electron density of the C-terminal region before the 9 end residues were added. (B) the electron density of the final C terminus. The protein is showed as stick and coloured in yellow for the structure before residues were added, and in green for the final structure. The $2F_o$-F_c map is contoured at 1σ and coloured in blue. The F_o-F_c is contoured at 3 σ and coloured in green for positive electron density and in red for negative electron density. Residues are noted, in 1 letter abbreviation, to show the differences between two structures. Picture created using Pymol.

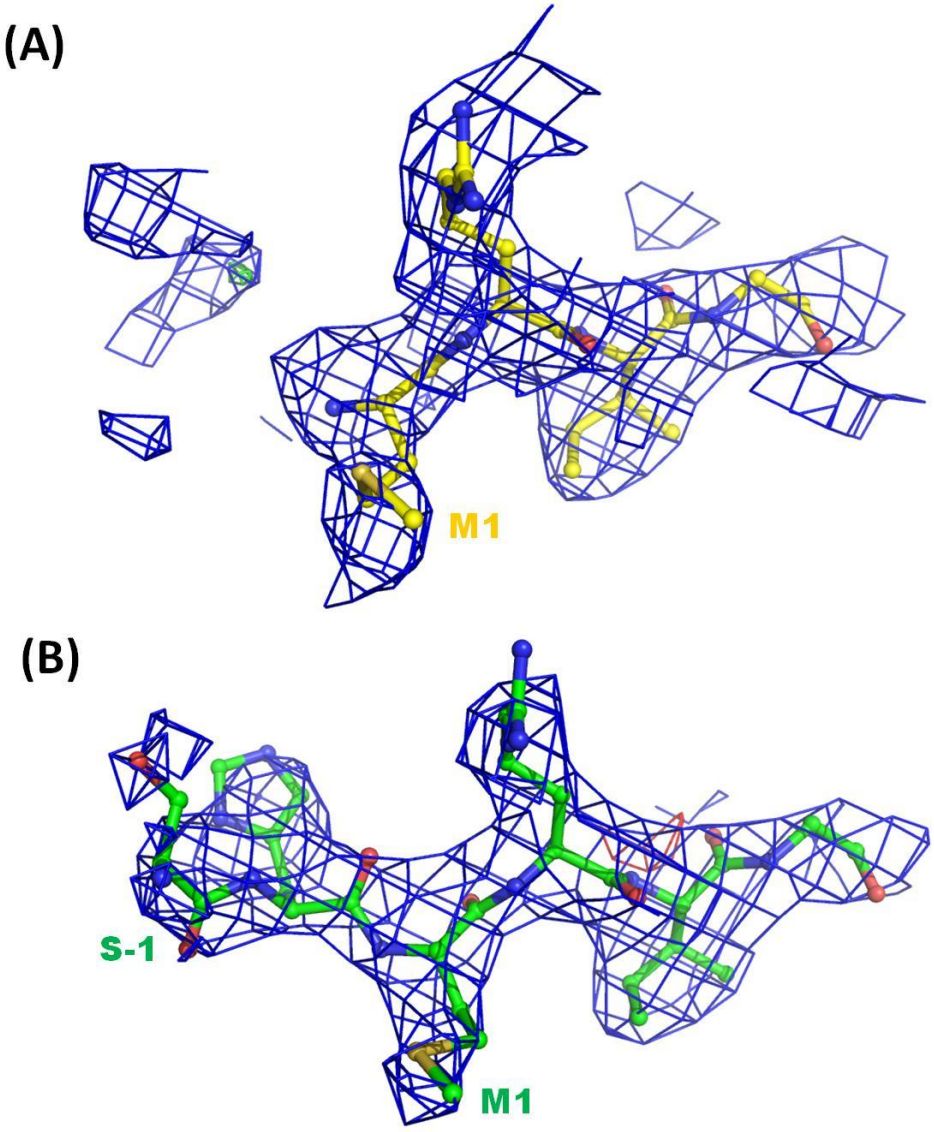


Figure 40. The electron density maps of the N terminus (of chain A as representative) of the structure BoGT6a in complex with UDP-GalNAc (derived from dataset 2) before and after the missing residues were built. (A) the electron density of the N-terminus before His0 and Ser-1 were added. (B) the electron density of the final N terminus. The protein is showed as stick and coloured in yellow for the structure before residues were added, and in green for the final structure. The $2F_o$-F_c map is contoured at 1σ and coloured in blue. The F_o-F_c is contoured at 3 σ and coloured in green for positive electron density and in red for negative electron density. Residues are noted, in 1 letter abbreviation, to show the differences between two structures. Picture created using Pymol.

The electron densities in the active sites had three different configurations. The first configuration, configuration A, was a big continuous component seen in chains E, F, G, H, O and P (Figure 41). The second, found in chains A, B, C, D, I, J, K and L, included two smaller components, the distance between which was comparable to the size of each component. This was called configuration B (Figure 42). The last configuration also contained two smaller components, but they were closer together. This configuration, configuration C, was found in chains M and N (Figure 43).

The whole UDP-GalNAc moieties were placed into all active sites according to the electron density. However, after refinement, UDP-GalNAc only fitted in electron density appearing in the active site of chain E, F, G, H, O and P (Figure 41B and C). Negative electron densities appeared in chain A, B, C, D, I, J, K, L, M and N where UDP-GalNAc had been added into the separated electron densities (Figure 42B, Figure 43B). UDP-GalNAc moieties of those chains were hence replaced with separate UDP and α-GalNAc moieties in chain A, B, C, D, I, J, K and L (Figure 42C). Although the resolution of the electron density map was not good enough to distinguish between α configuration and β configuration, the GalNAc moieties were set as α-GalNAc (noted as GalNAc) because BoGT6a is a retaining glycosyltransferase and so its product must be in the same configuration as that of the GalNAc in UDP-GalNAc.

Interestingly, in the active sites of chain M and chain N that had electron density in form C, the GalNAc moieties appeared to be in close proximity, about 1.3 Å, to residue Gln192. The oxygen atoms of the hydroxyl groups of the C1 atoms of the GalNAc moieties were deleted and links between the nitrogen atoms of the amino group and the C1 were created. GalNAc moieties appeared in β configuration (abbreviated as NGA) when they linked to Gln192, and in α configuration when they were free in the active site (Figure 43C).

After completing the missing C-terminal regions, the N-terminal regions and the substrates in the active sites, there was still some positive difference electron density remaining around the chains. Some glycerol moieties, some water molecules and one SO_4^{2-} ion were added to the structure (Figure 44, Figure 45, Figure 46).

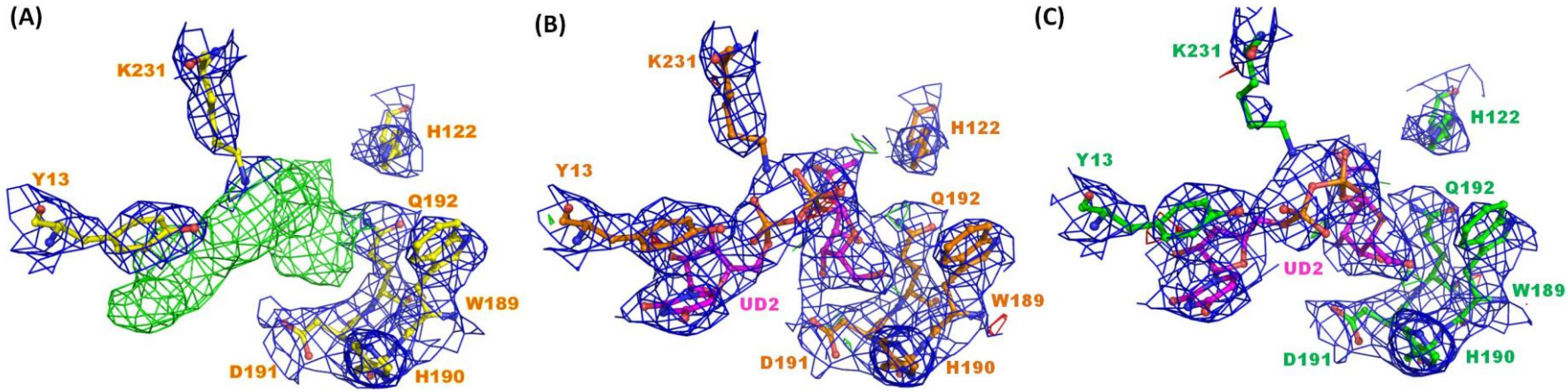


Figure 41. The conformation A of the electron densities that appeared in the active sites of the complex BoGT6a E192Q•UDP-GalNAc (derived from dataset 2). (A) shows the positive difference electron densities that appeared in the active site of chain E (as representative). (B) shows the electron density map of the active site of chain E resulted from the first refinement round after UDP-GalNAc (short as UD2) was added in to the structure. (C) shows the electron density map of the active site of chain E of the final structure. The protein is showed as stick and coloured in yellow for the structure before UD2 was added, in orange when UD2 was added and in green for the final structure. The $2F_o$-F_c map is contoured at 1σ and coloured in blue. The F_o-F_c is contoured at 3 σ and coloured in green for positive electron density and in red for negative electron density. Residues are noted, in 1 letter abbreviation, to indicate the active site of the enzyme. Picture created using Pymol.

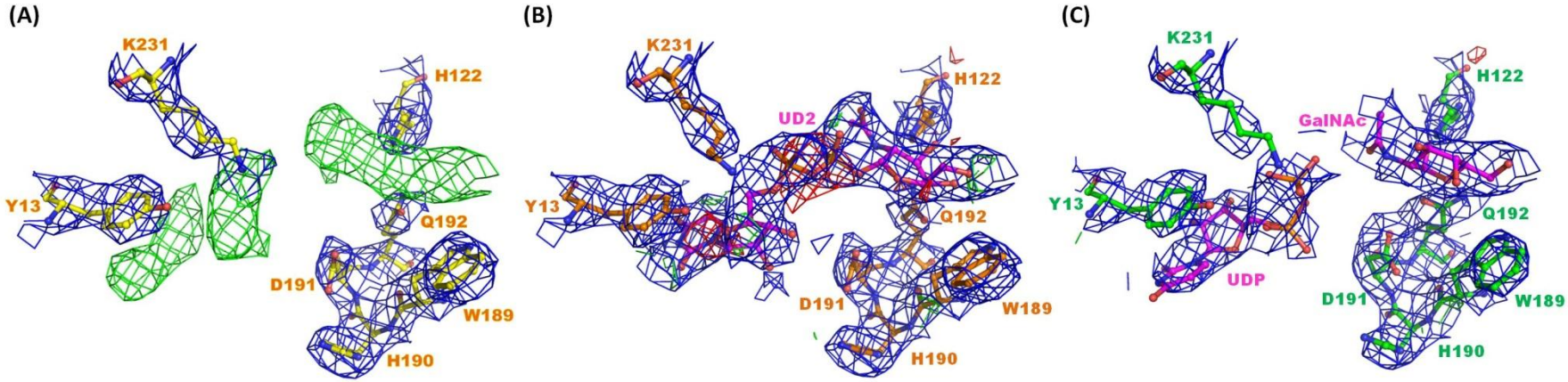


Figure 42. The conformation B of the electron densities that appeared in the active sites of the complex BoGT6a E192Q•UDP-GalNAc (derived from dataset 2). (A) shows the positive difference electron densities appearing in the active sites of chain A (as representative). (B) shows the electron density map of the active site of chain A resulted from the first refinement round after UDP-GalNAc (short as UD2) was added in to the structure. Negative electron density appeared between UDP and GalNAc moiety. (C) shows the electron density map of the active site of chain A of the final structure which the UD2 was replaced by separated UDP and α-GalNAc (noted as GalNAc). The protein is showed as stick and coloured in yellow for the structure before UD2 was added, in orange when UD2 was added and in green for the final structure. The $2F_o$-F_c map is contoured at 1σ and coloured in blue. The F_o-F_c is contoured at 3 σ and coloured in green for positive electron density and in red for negative electron density. Residues are noted, in 1 letter abbreviation, to indicate the active site of the enzyme. Picture created using Pymol.

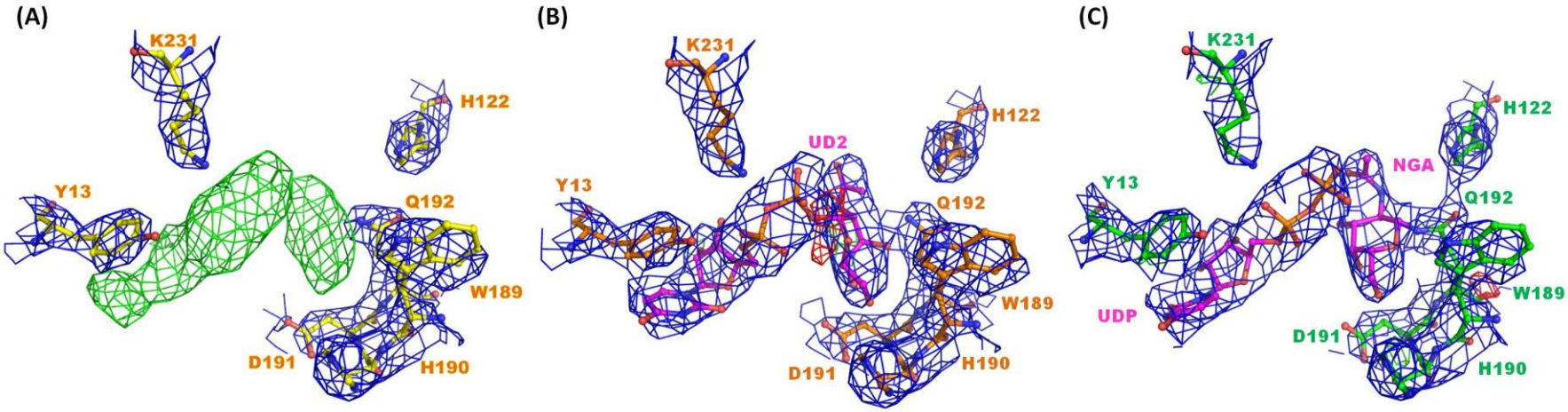


Figure 43. The conformation C of the electron densities that appeared in the active sites of the complex BoGT6a E192Q•UDP-GalNAc (derived from dataset 2). (A) shows the positive difference electron densities appearing in the active sites of chain M (as representative). (B) shows the electron density map of the active site of chain M resulted from the first refinement round after the UDP-GalNAc (short as UD2) moieties were added in to the structure. Negative electron density appeared at the bonds between UDP and GalNAc moieties. (D) shows the electron density map of the active site of chain M of the final structure which UDP-GalNAc was replaced by separated UDP and β-GalNAc (noted as NGA). The protein is showed as stick and coloured in yellow for the structure before UD2 was added, in orange when UD2 was added and in green for the final structure. The $2F_o$-F_c map is contoured at 1σ and coloured in blue. The F_o-F_c is contoured at 3 σ and coloured in green for positive electron density and in red for negative electron density. Residues are noted, in 1 letter abbreviation, to indicate the active site of the enzyme. Picture created using Pymol.

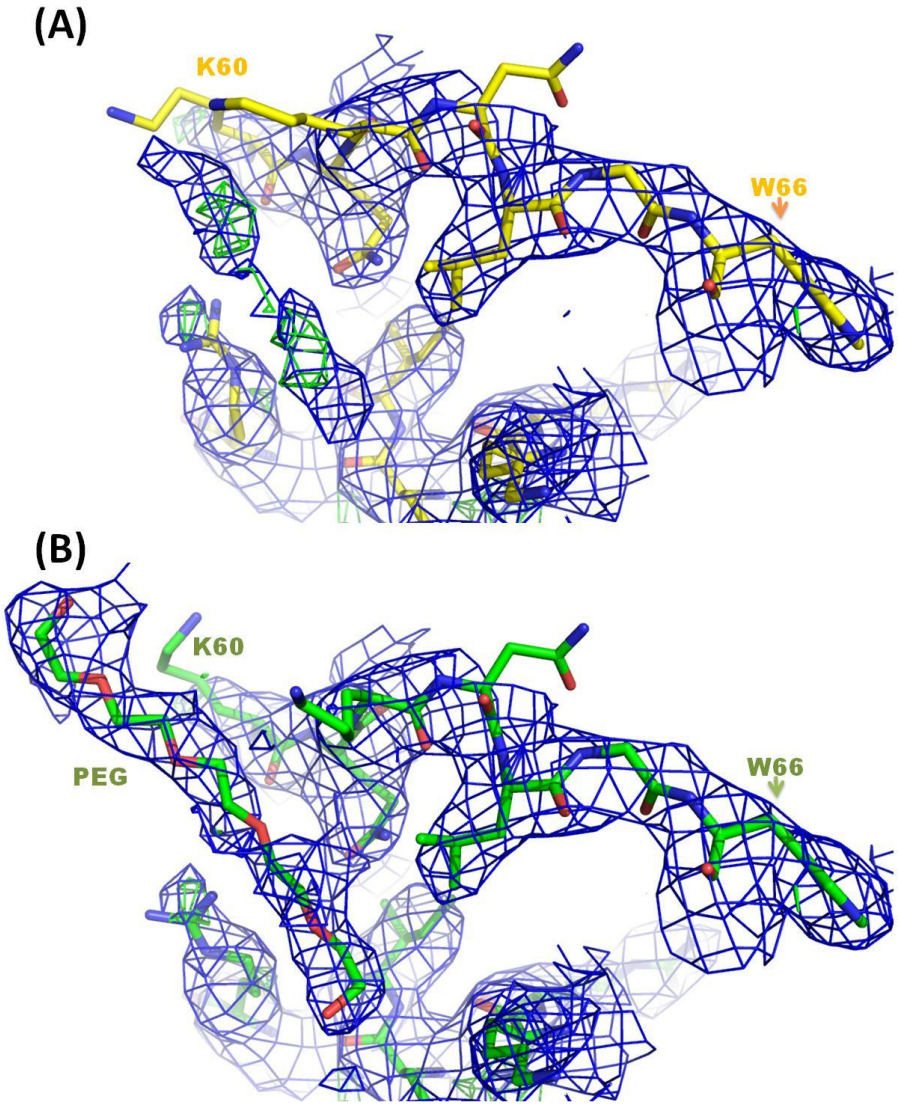


Figure 44. Electron density of PEG in chain A of the structure BoGT6a E192Q in complex with UDP-GalNAc derived from the dataset 2. (A) shows electron density map before PEG was added to the structure. (B) shows electron density map of the final structure which PEG was added to the structure. Protein is shown as line in yellow for the structure before PEG was added and in green for the final structure. The $2F_o$-F_c map is contoured at 1σ and coloured in blue. The F_o-F_c is contoured at 3 σ and coloured in green for positive electron density and in red for negative electron density. Residues are noted, in 1 letter abbreviation, to indicate the location of PEG. Picture created using Pymol.

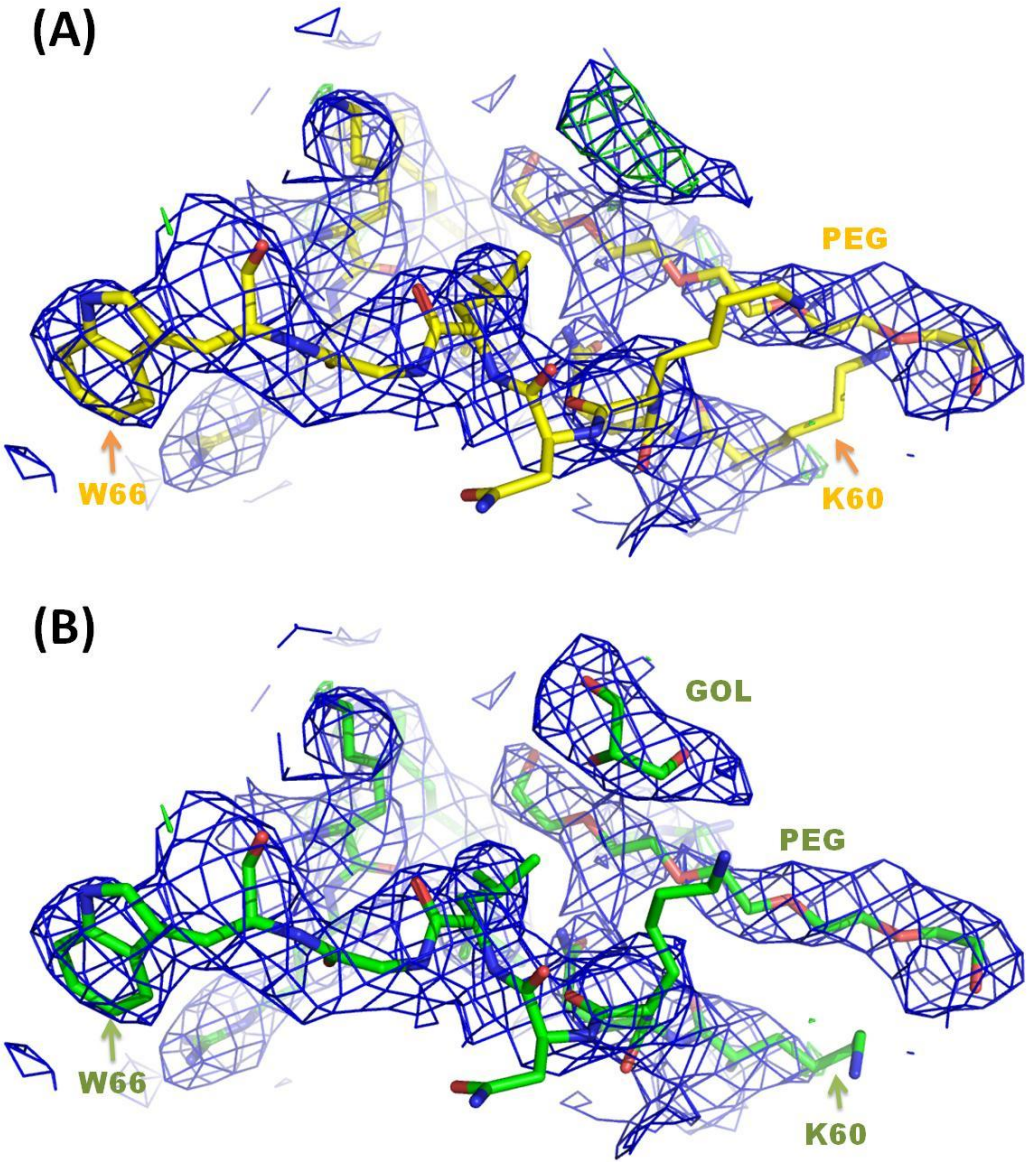


Figure 45. Electron density of glycerol in chain A (as representative) of the structure BoGT6a E192Q in complex with UDP-GalNAc derived from the dataset 2. (A) shows electron density map before glycerol (noted as GOL) was added to the structure. (B) shows electron density map of the final structure which GOL was added to the structure. Protein is shown as line in yellow for the structure before PEG was added and in green for the final structure. The $2F_o$-F_c map is contoured at 1σ and coloured in blue. The F_o-F_c is contoured at 3 σ and coloured in green for positive electron density and in red for negative electron density. Residues are noted, in 1 letter abbreviation, to indicate the location of GOL. Picture created using Pymol.

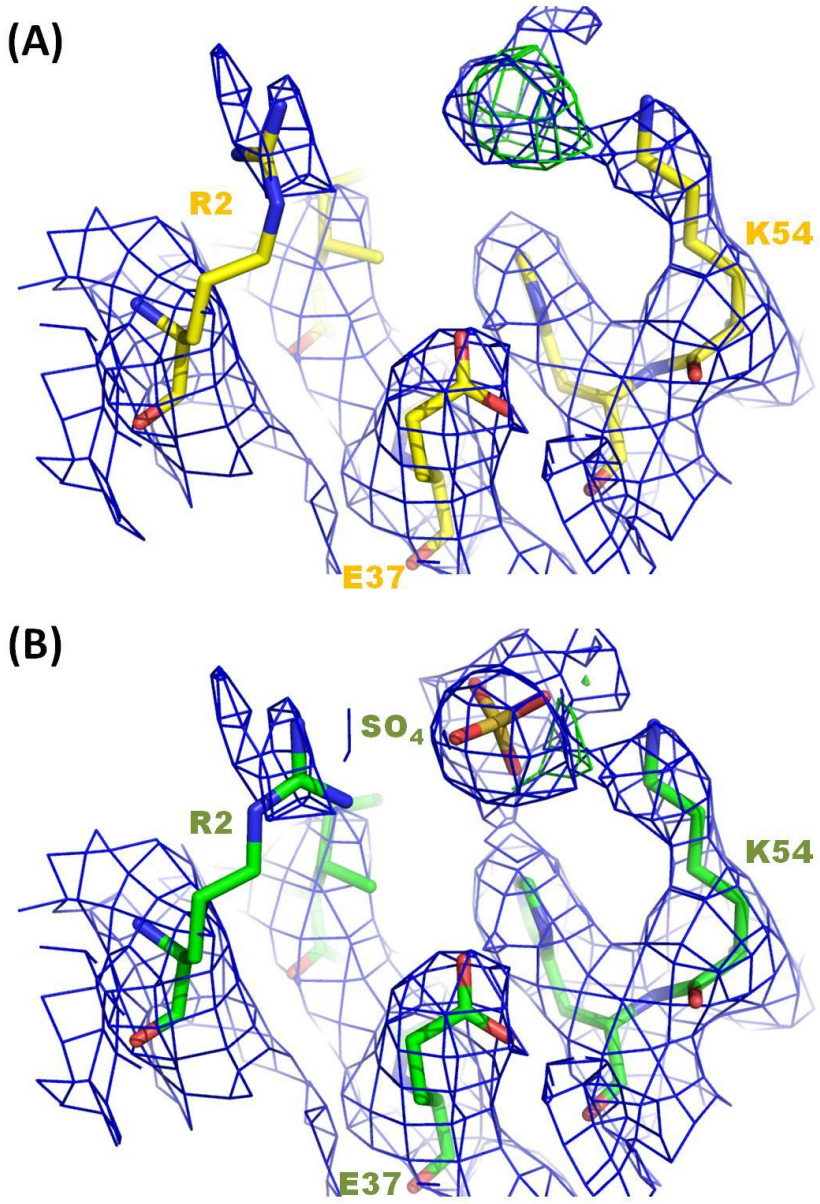


Figure 46. Electron density of SO_4^{2-} ion in chain B of the structure BoGT6a E192Q in complex with UDP-GalNAc derived from the dataset 2. (A) shows electron density map before SO_4^{2-} ion (noted as SO_4) was added to the structure. (B) shows electron density map of the final structure which SO_4^{2-} was added to the structure. Protein is shown as line in yellow for the structure before SO_4^{2-} ion was added and in green for the final structure. The $2F_o$-F_c map is contoured at 1σ and coloured in blue. The F_o-F_c is contoured at 3 σ and coloured in green for positive electron density and in red for negative electron density. Residues are noted, in 1 letter abbreviation, to indicate the location of SO_4^{2-} ion. Picture created using Pymol.

Improving Ramachandran values and fitting residues into the electron density was performed during a few refinement cycles until the R/Rfree values reached 22.53/24.94 and 93.6 % of residues were in the favoured region of the Ramachandran plot. More crystallographic statistics are shown in Table 7. The final structure was named BoGT6a E192Q•UDP-GalNAc structure in monoclinic form with 16 molecules in the asymmetric unit (Figure 47). Of these, 6 chains (E, F, G, H, O and P) had UDP-GalNAc (configuration A) in their active sites. 8 chains (A, B, C, D, I, J, K and L) had UDP and α-GalNAc (configuration B) in their active sites, whilst 2 chains (M and N) had UDP and β-GalNAc (configuration C) in their active sites (Figure 47, Figure 48).

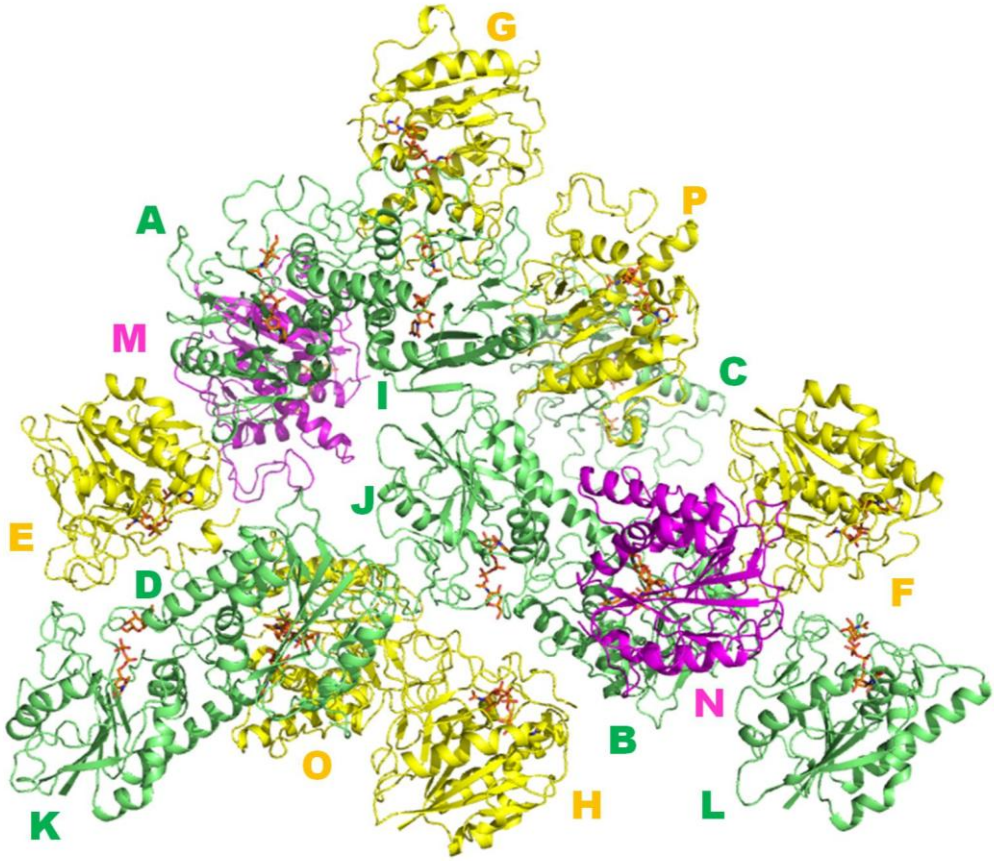


Figure 47. Crystal structure of the BoGT6a E192Q in complex with the donor UDP-GalNAc from the monoclinic crystal form (form III). The protein is shown in cartoon representation. The molecules with intact UDP-GalNAc are coloured in yellow, the molecules with UDP-GalNAc and α-GalNAc in green, and the molecules with UDP-GalNAc and β-GalNAc in magenta. The ligands are shown as orange sticks. Each chain is labelled as their name. Picture created using Pymol.

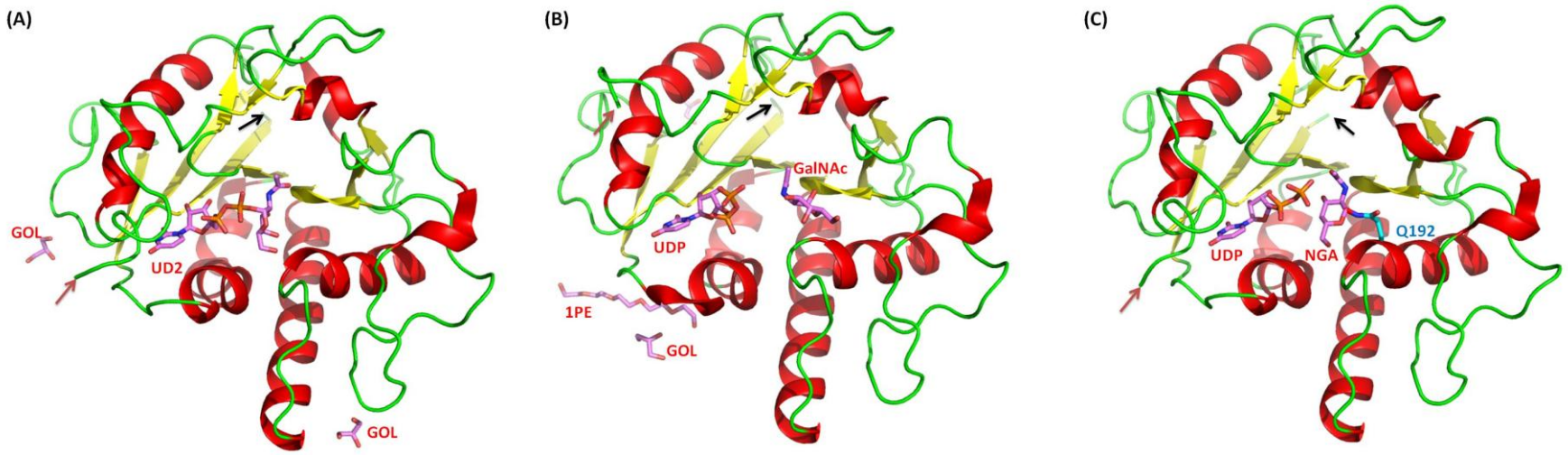


Figure 48. Comparison of the overall structure of representative molecules in the BoGT6a E192Q in complex with the donor UDP-GalNAc form III structure. (A) shows the molecule with intact UDP-GalNAc. (B) shows the molecule with UDP and α-GalNAc. (C) shows the molecule with UDP and β-GalNAc. Protein is shown in cartoon representation and coloured by secondary structure. The ligand is shown as stick in violet and labelled. UDP-GalNAc is noted as UD2, α-GalNAc as GalNAc and β-GalNAc as NGA. Red arrows indicate the C-termini, and black arrows the N-termini. Picture created using Pymol.

3.3.2.3 Structure of BoGT6a in complex with UDP-GalNAc (form II)

As was the case for dataset 5, dataset 4 was processed automatically by XIA2 program in $P2_12_12_1$, but with lower resolution (3.42 Å). The cell content analysis result from Matthews_Coef program in CCP4i indicated that there were also 4 molecules per asymmetric unit with a solvent content of 55.06 %. MR using Phaser program in PHENIX with chain BoGT6a•FAL without FAL moiety as a starting model found 4 molecules after one search. The values of LLG and TFZ were 2597 and 13.8 respectively.

Since the quality of this dataset was lower than that of dataset 5, dataset 5 was given priority for processing. When the structure of BoGT6a E192Q•GalNAc was solved, dataset 4 was analysed again, using BoGT6a E192Q•GalNAc without GalNAc as a starting model. The searching experiment also gave a solution with 4 molecules per asymmetric unit, but the values of LLG and TFZ were slightly different. Although the TFZ value was lower (9.6 compared to 13.8), the solution from this search was used for further analysis because the LLG was significantly higher (2764 compared to 2597) and the resulting structure had a longer C-terminal region (terminating at Tyr236 instead of Lys231 as in BoGT6a•FAL) (Figure 49). The first refinement result also gave slightly improved values of R and R_{free}, 28.08 and 37.65 respectively, compared to 29.32 and 38.95 from the first refinement with the previous Phaser result using BoGT6a•FAL without the FAL moiety as the searching model.

Like the starting map of the other structures, there was positive difference electron density in the active sites and the C-terminal regions of all chains. At the C-termini, there was also a difference in the length of each chain. Chains A and C had larger difference electron densities than those of the other chains. After missing residues were added into the structure using Coot and several rounds of refinement were performed using Refine in PHENIX, the final structure consisted of 241 residues in chain A, 238 residues in chain B, 240 residues in chain C, and 236 residues in chain D (Figure 49, Figure 50).

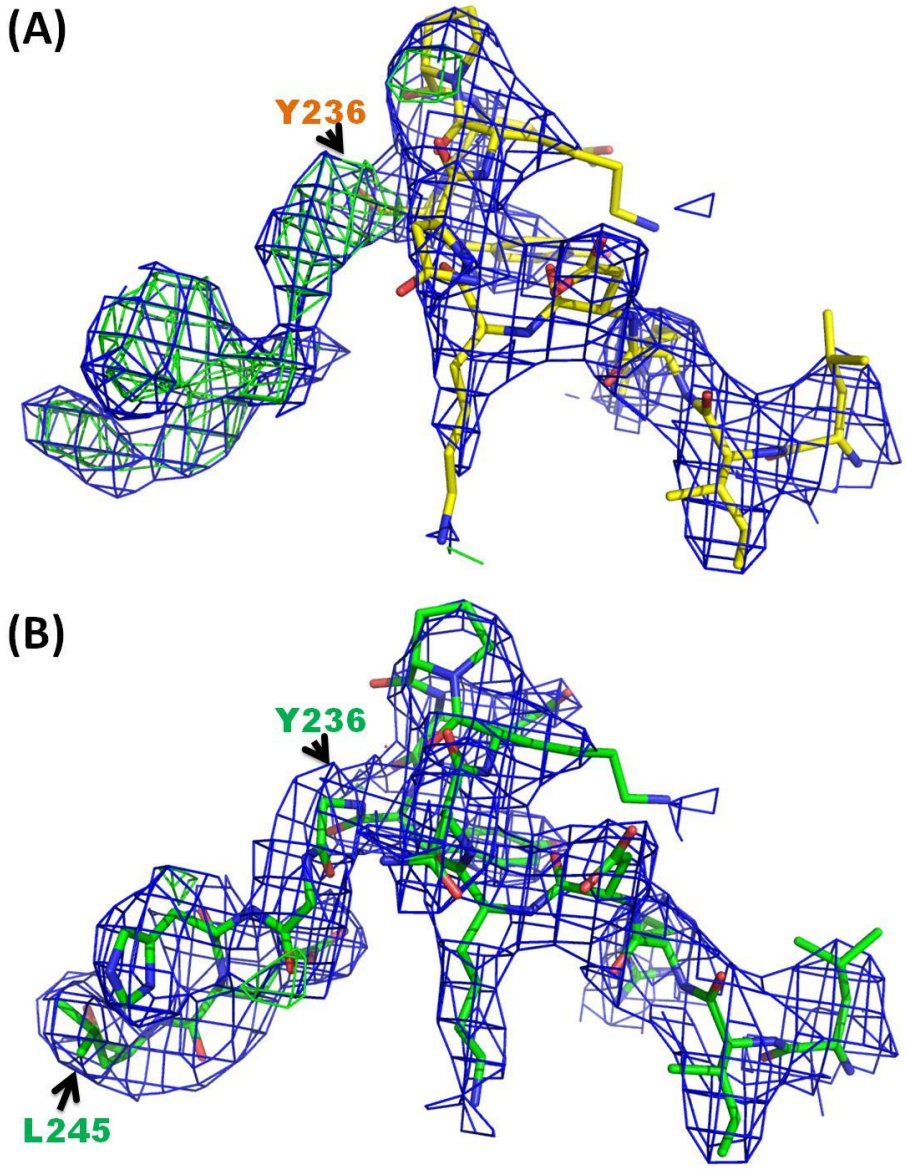


Figure 49. Electron density map of a long C terminus in the structure of BoGT6a E192Q•UDP-GalNAc derived from the dataset 4. (A) shows the positive difference electron density map of the C terminal region of chain A (as representative) after using MR method with the structure of BoGT6a E192Q•GalNAc without GalNAc as a searching model. (B) shows the electron density map of the C terminus of the final structure. The $2F_o$-F_c map is contoured at 1σ and coloured in blue. The F_o-F_c is contoured at 3 σ and coloured in green for positive electron density and in red for negative electron density. Residues Tyr236 and Leu245 are noted, in 1 letter abbreviation, to indicate the difference between two structures. Picture created using Pymol.

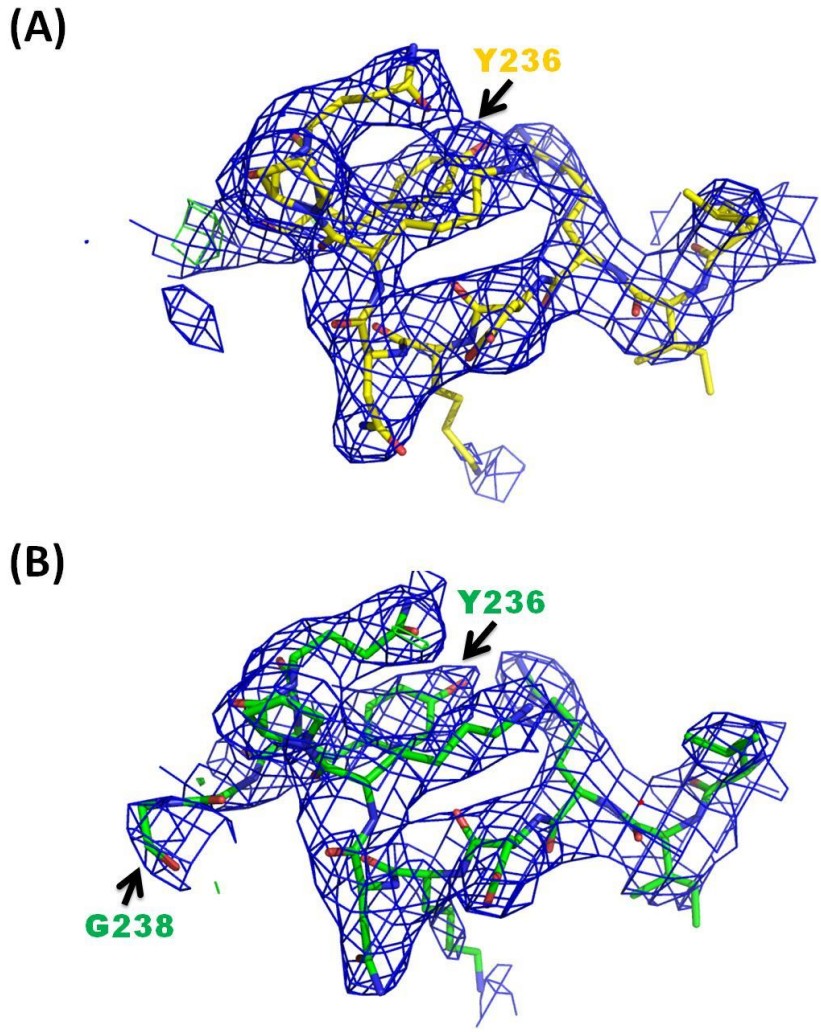


Figure 50. Electron density map of a short C terminus in the structure of BoGT6a E192Q•UDP-GalNAc derived from the dataset 4. (A) shows the positive difference electron density map of the C terminal region of chain B (as representative) after using MR method with the structure of BoGT6a E192Q•GalNAc without GalNAc as a searching model. (B) shows the electron density map of the C terminus of the final structure. The $2F_o$-F_c map is contoured at 1σ and coloured in blue. The F_o-F_c is contoured at 3 σ and coloured in green for positive electron density and in red for negative electron density. Residues Tyr236 and Gly238 are noted, in 1 letter abbreviation, to indicate the difference between two structures. Picture created using Pymol.

Unlike the structure of BoGT6a E192Q in complex with UDP-GalNAc in monoclinic form, the positive densities in the active sites of this structure had only two of the three configurations found in the previous donor bound BoGT6a E192Q map. Configuration A was seen in chain A and chain C and configuration B in chain

B and chain D. UDP-GalNAc moieties were added into the chain A and chain C active sites, and UDP and α-GalNAc moieties were added into chain B and chain D (Figure 51, Figure 52).

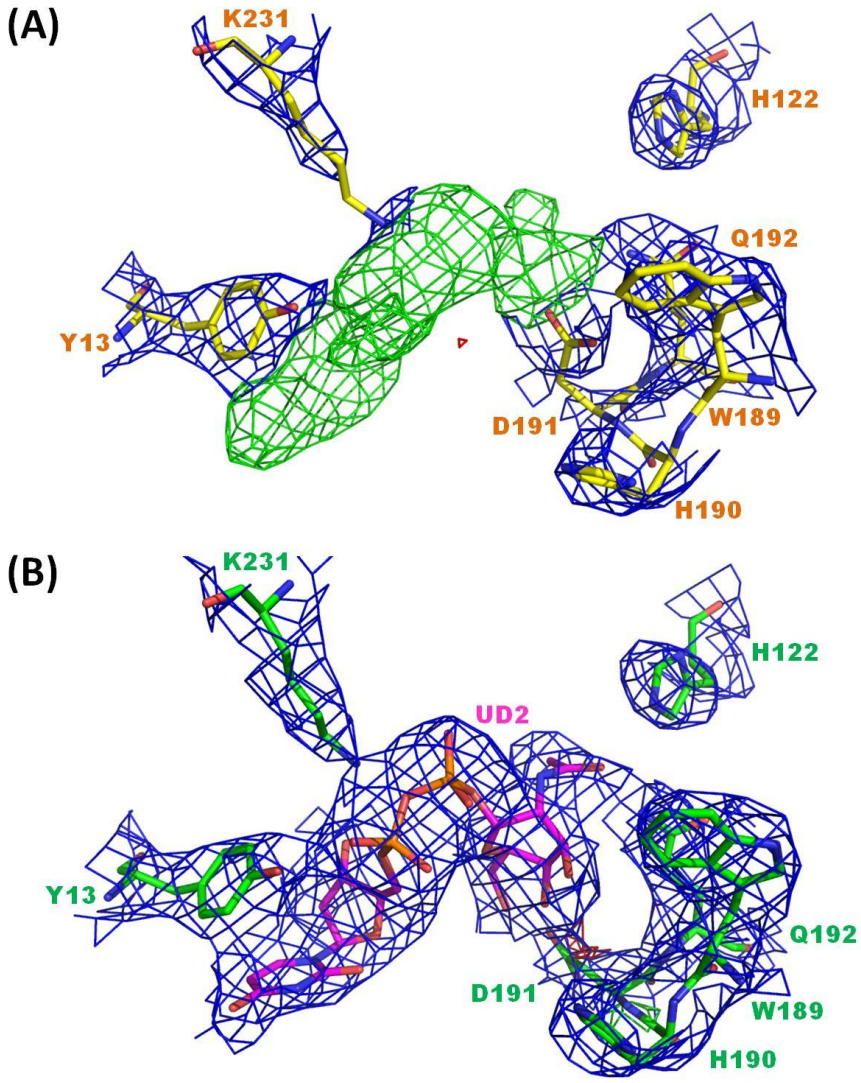


Figure 51. The conformation A of the electron densities that appeared in the active sites of the complex BoGT6a E192Q•UDP-GalNAc (derived from dataset 4). (A) shows the positive electron densities appearing in the active site of chain A (as representative). (B) shows the electron density map of the active site of chain A of the final structure. The protein is showed as line and coloured in yellow for the structure before UDP-GalNAc (noted as UD2) was added and in green for the final structure. The $2F_o$-F_c map is contoured at 1σ and coloured in blue. The F_o-F_c is contoured at 3 σ and coloured in green for positive electron density and in red for negative electron density. Residues are noted, in 1 letter abbreviation, to indicate the active site of the enzyme. Picture created using Pymol.

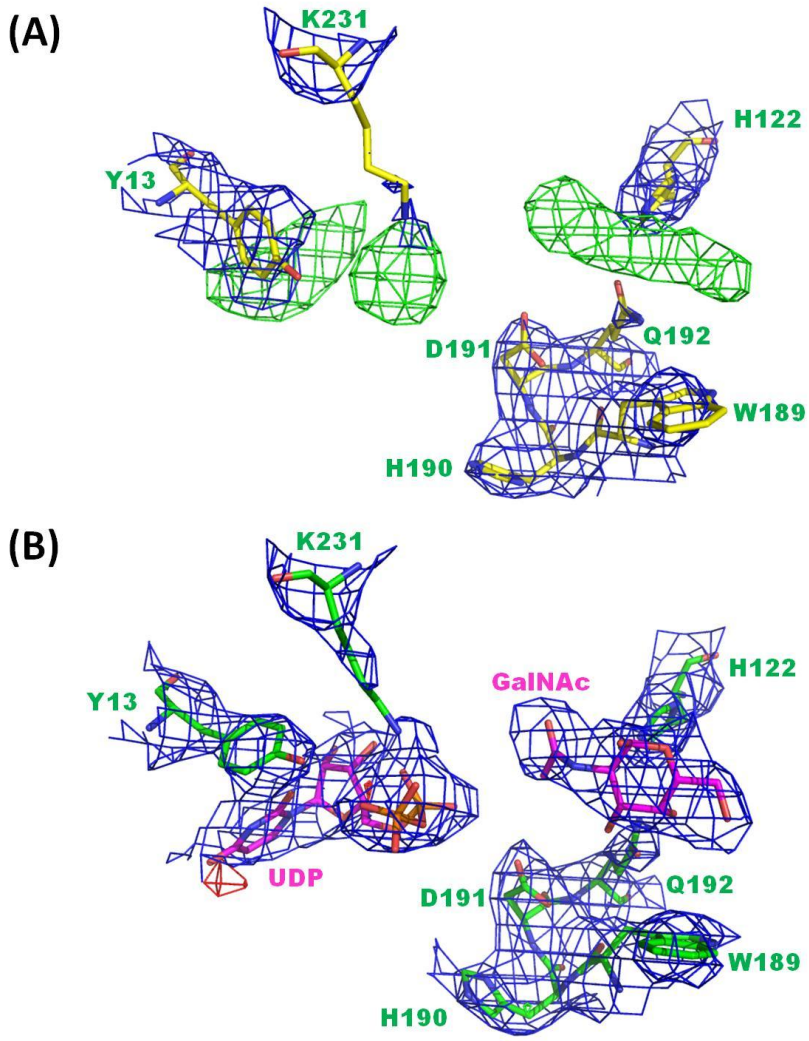


Figure 52. The conformation B of the electron densities that appeared in the active sites of the complex BoGT6a E192Q•UDP-GalNAc (derived from dataset 4). (A) shows the positive electron densities appearing in the active site of chain B (as representative). (B) shows the electron density map of the active site of chain B of the final structure with UDP and α-GalNAc (noted as GalNAc). The protein is showed as line and coloured in yellow for the structure without ligands and in green for the final structure. The $2F_o$-F_c map is contoured at 1σ and coloured in blue. The F_o-F_c is contoured at 3 σ and coloured in green for positive electron density and in red for negative electron density. Residues are noted, in 1 letter abbreviation, to indicate the active site of the enzyme. Picture created using Pymol.

After several cycles of refinement, the final values of R and R_{free} were 28.35 and 31.41 respectively and 92.7 % of residues were in the favoured region of the Ramachandran plot. Other crystallographic statistics are listed in Table 7. The final

structure was called BoGT6a E192Q•UDP-GalNAc structure in orthorhombic form with 4 molecules per asymmetric unit, with 2 chains (A and C) having UDP-GalNAc in their active sites and 2 chains (B and D) having UDP and α-GalNAc (Figure 53).

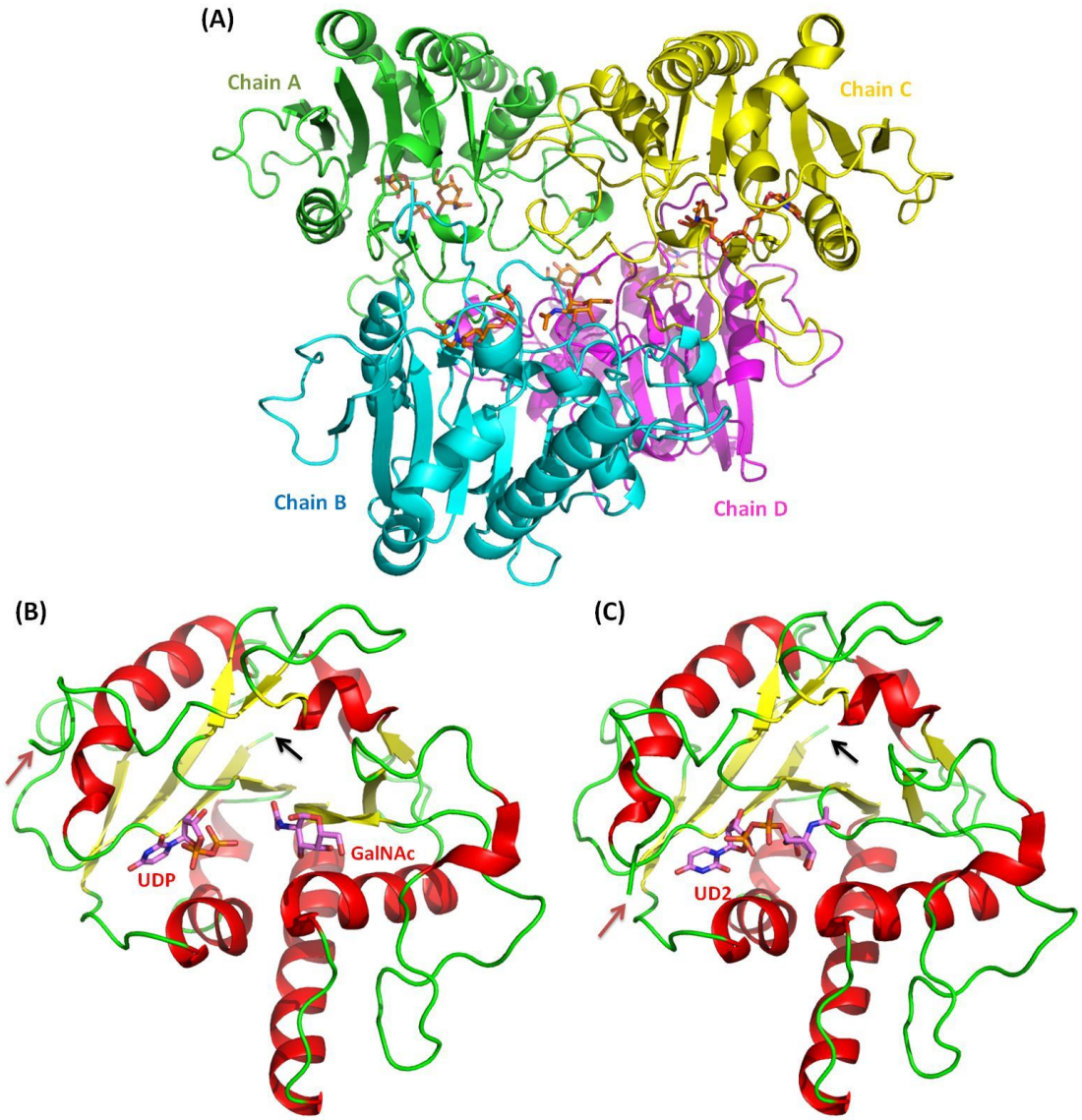


Figure 53. Crystal structure of BoGT6a E192Q in complex with the donor UDP-GalNAc in orthorhombic crystal form (form II). (A) 4 chains in a asymmetric unit. Protein is shown in cartoon representation and coloured by chain. The ligands were shown as orange sticks. (B) shows chain A and (C) chain B as representatives. Protein is shown in cartoon representation and coloured by secondary structure. The ligands are shown as stick in violet and labelled. UDP-GalNAc is noted as UD2, α-GalNAc as GalNAc. Red arrow indicates the C-terminus, and black arrow the N-terminus. Picture created using Pymol.

Table 7. X-ray crystallographic statistics

	BoGT6a E192Q•GalNAc	**BoGT6a E192Q•UDP-GalNAc**	**BoGT6a E192Q•UDP-GalNAc**
Ligands used in crystallisation	UDP-GalNAc	UDP-GalNAc	UDP-GalNAc
Ligands observed in crystal structure	GalNAc	UDP-GalNAc, UDP, GalNAc	UDP-GalNAc, UDP, GalNAc
Space Group	$P\ 2_12_12_1$	$P\ 2_12_12_1$	$P\ 2_1$
No. of molecules/a.u	4	4	16
Cell dimensions	a= 80.1, b= 115.6, c= 126.1 Å	a= 80.1, b= 120.1, c= 131.8 Å	a=177.0, b= 79.8, c= 179.1 Å, β= 95.2°
Resolution range (Å)	67.6 –2.8 (2.9 – 2.8)	88.8 –3.4 (3.6 – 3.4)	88.0 –3.5 (3.6 – 3.5)
R_{merge} (outer shell)	0.10 (0.71)	0.09 (0.54)	0.13 (0.50)
I/σI (outer shell)	15.0 (2.3)	13.5 (2.8)	7.4 (2.1)
Completeness (outer shell) %	97.7 (99.5)	97.6 (99.8)	98 (99.0)
Total no. of reflections	158394	93516	168581
Unique no. of reflections	29475	17402	61949
Redundancy (outer shell)	5.4 (4.9)	5.4 (4.6)	2.7 (2.6)
Wilson B-factor ($Å^2$)	45.73	93.61	76.41
R/R_{free}	23.14/27.35	28.35/31.41	22.53/24.94
Overall average B-factor ($Å^2$)	41.19	82.25	71.38
Number of Protein chains UDP-GalNAc UDP GalNAc Water PEG SO_4^{2-} Glycerol	4 - - 4 174 - - -	4 2 2 2 - - - -	16 6 10 10 7 1 1 5
RMSD values bond length (Å) bond angle (^{o})	 0.004 0.979	 0.006 1.206	 0.002 0.529
Ramachandran plot statistics (%) Favoured Outliers	 96.15 0.11	 92.71 0.21	 93.66 0.10
RCSB-PDB codes	**4cjb**	**4cjc**	**4cj8**

3.3.3 Discussion

Although both BoGT6a native form and BoGT6a E192Q were set up with the donor substrate, UDP-GalNAc, only the BoGT6a E192Q complex formed crystals. This could be as a result of the slow catalytic activity of the mutant (22000 fold reduction compared to the native enzyme). All bovine α3GT and human GTA/GTB are reported to have high catalytic rates, meaning that when the native enzymes were mixed with their donor substrate and/or the acceptor substrates, the glycosyl transfer or the hydrolysis always happened during the crystallisations, destabilising the proteins and, as a result, inhibiting the crystal formation. There has been no published structure of native GT6 with the intact donor substrate UDP-Gal or UDP-GalNAc. There are three forms of BoGT6a E192Q in complex with UDP-GalNAc structures, including two orthorhombic forms and one monoclinic form.

Both orthorhombic forms have 4 molecules in the asymmetric unit, but the substrates in their active site are different. Although UDP-GalNAc was present in the crystallisation solution, in the form I structure, only GalNAc was observed in the active site of all 4 molecules (Figure 54A). In contrast, in the form II structure, two molecules (chains A and C) contain intact UDP-GalNAc and the others (chains B and D) contain the hydrolysis products, UDP and α-GalNAc (noted as GalNAc) (Figure 54B and C). This shows that the BoGT6a E192Q mutant still retains sufficient hydrolysis activity to catalyse the reaction during the crystallisation. Form II crystals were obtained from the complex that had been incubated for 1 h, whereas the form I crystals that contain only GalNAc were grown from a preparation that had been incubated overnight, suggesting that the mutant GT had hydrolysed most of the substrate and the UDP product had dissociated from the protein molecules in the crystals.

The monoclinic form structure (form III), was also prepared with protein that had been pre-incubated for only 1 h, but it is less ordered than the form II structure and so the crystal diffracted to a lower resolution. Interestingly, it consists of 16 polypeptide chains in the asymmetric unit, which contain three different ligand configurations, designated as configuration A, B and C. Although the structure was

solved at a low resolution (3.5 Å), the quality of the electron density map and clearly showed the electron densities of the ligands (Figure 54D, E and F).

Configuration A, found in chains E, F, G, H, O and P, is an intact UDP-GalNAc bound in a compact conformation similar to chains A and C of orthorhombic form II. Configuration B, seen in chains A, B, C, D, I, J, K and L, includes UDP and α-GalNAc. This is similar to the ligand conformation observed in chains B and D of the form II structure. The last configuration, C, found in chains M and N, also consists of two separate components, but the sugar moiety is close to the UDP and remains in a similar orientation and location to that in UDP-GalNAc. Due to the close distance between the sugar moiety and the residue Gln192, a link was created between them and the sugar moiety was set in β configuration (noted as NGA) (Figure 54F).

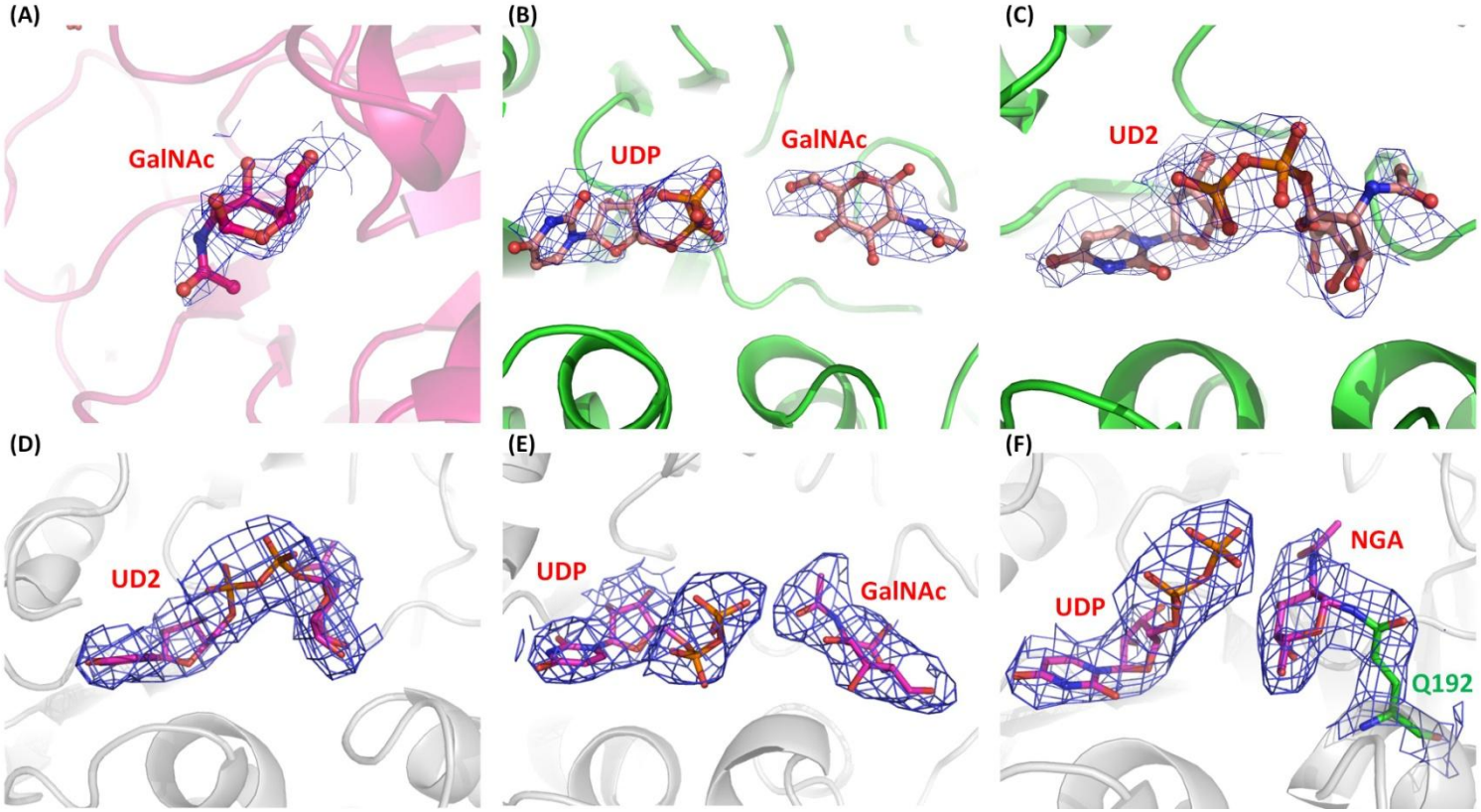


Figure 54. Electron densities of ligands in 3 structures of BoGT6a E192Q in complex with the donor UDP-GalNAc. (A) shows a ligand in chain A (as representative) of the form I structure. (B) and (C) show ligands in chain A and in chain B (as representative) of the form II structure respectively. (D), (E), and (F) show ligands in chain E, chain A, and chain M (as representative) of the form III structure respectively. Proteins are showed in cartoon representation. Ligands are shown as stick in magenta. α-GalNAc is noted as GalNAc, UDP-GalNAc as UD2, and β-GalNAc as NGA. The $2F_o$-F_c map is contoured at 1σ and coloured in blue. Residue Gln192 is noted, in 1 letter abbreviation, to show the bond. Picture created using Pymol.

3.3.3.1 Symmetries in the structures of BoGT6a in complex with UDP-GalNAc

After the three structures had been solved, the packing of the molecules in their asymmetric units attracted our attention, especially in the structure of BoGT6a E192Q•UDP-GalNAc in monoclinic form. In an attempt to understand the pattern of the packing style in the monoclinic complex structure, all the structures, including the BoGT6a•FAL and all mutant BoGT6a in complex with UDP-GalNAc, were compared to each other.

The form I structure, BoGT6a E192Q•GalNAc, had 4 molecules per asymmetric unit, but at the first look, the packing was clearly different from that of the structure BoGT6a•FAL. Meanwhile the BoGT6a E192Q•UDP-GalNAc in orthorhombic form had a similar packing arrangement to that of BoGT6a•FAL (Figure 55).

Attempting to superpose chain A of the BoGT6a E192Q•GalNAc structure with each chain of the BoGT6a•FAL structure illustrated that the positions of the chains in its asymmetric unit were different from those of the chains in the other structure. Nevertheless, when the symmetrical coordinates of the BoGT6a E192Q•GalNAc structure were shown, chain A of the BoGT6a E192Q•GalNAc structure superposed with chain D of BoGT6a E192Q•GalNAc structure, and its other chains superposed with the symmetrical coordinates (Figure 56B). The organisation of the chains of the BoGT6a E192Q•GalNAc structure was rearranged according to that of the BoGT6a•FAL structure (Figure 56C). This indicated that the relationship between the new positions of chains A, B and C and their old positions was pseudo translation. The new structure of BoGT6a E192Q•GalNAc also contains 4 molecules per asymmetric unit, with the position of the β-GalNAc similar to the position of FAL in the BoGT6a•FAL structure.

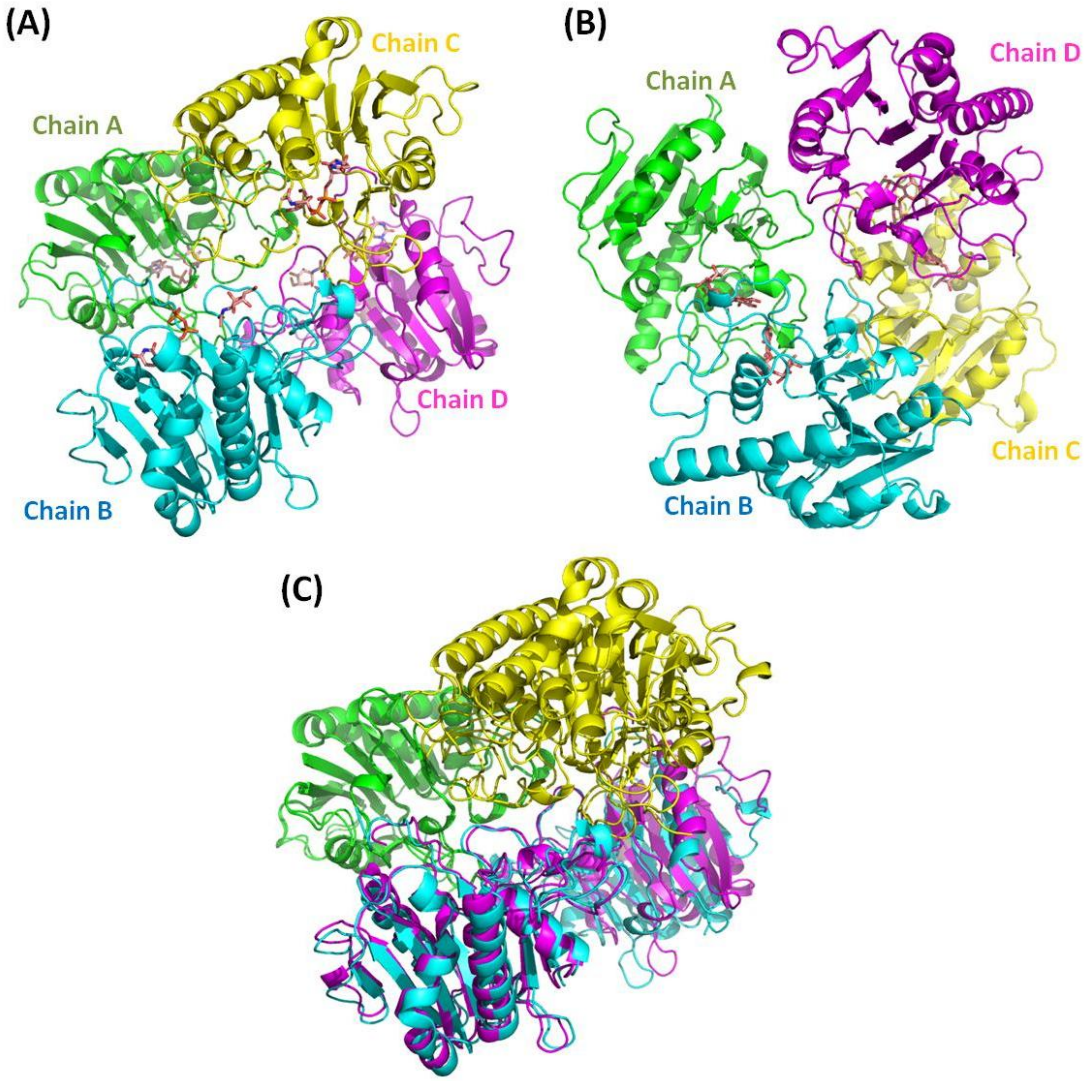


Figure 55. Comparison of the arrangement of molecules in the asymmetric unit of the BoGT6a E192Q•UDP-GalNAc structure in form II and that of the BoGT6a•FAL structure. (A) 4 molecules in the asymmetric unit of the BoGT6a E192Q•UDP-GalNAc form II structure. (B) 4 molecules in asymmetric unit of the BoGT6a•FAL structure. (C) a superposition of chain A of the BoGT6a E192Q•UDP-GalNAc form II structure onto chain A of the BoGT6a•FAL structure. Proteins are shown in cartoon representation where chain A is coloured in green, chain B in cyan, chain C in yellow, and chain D in magenta. The substrates are shown as orange sticks. Picture created using Pymol.

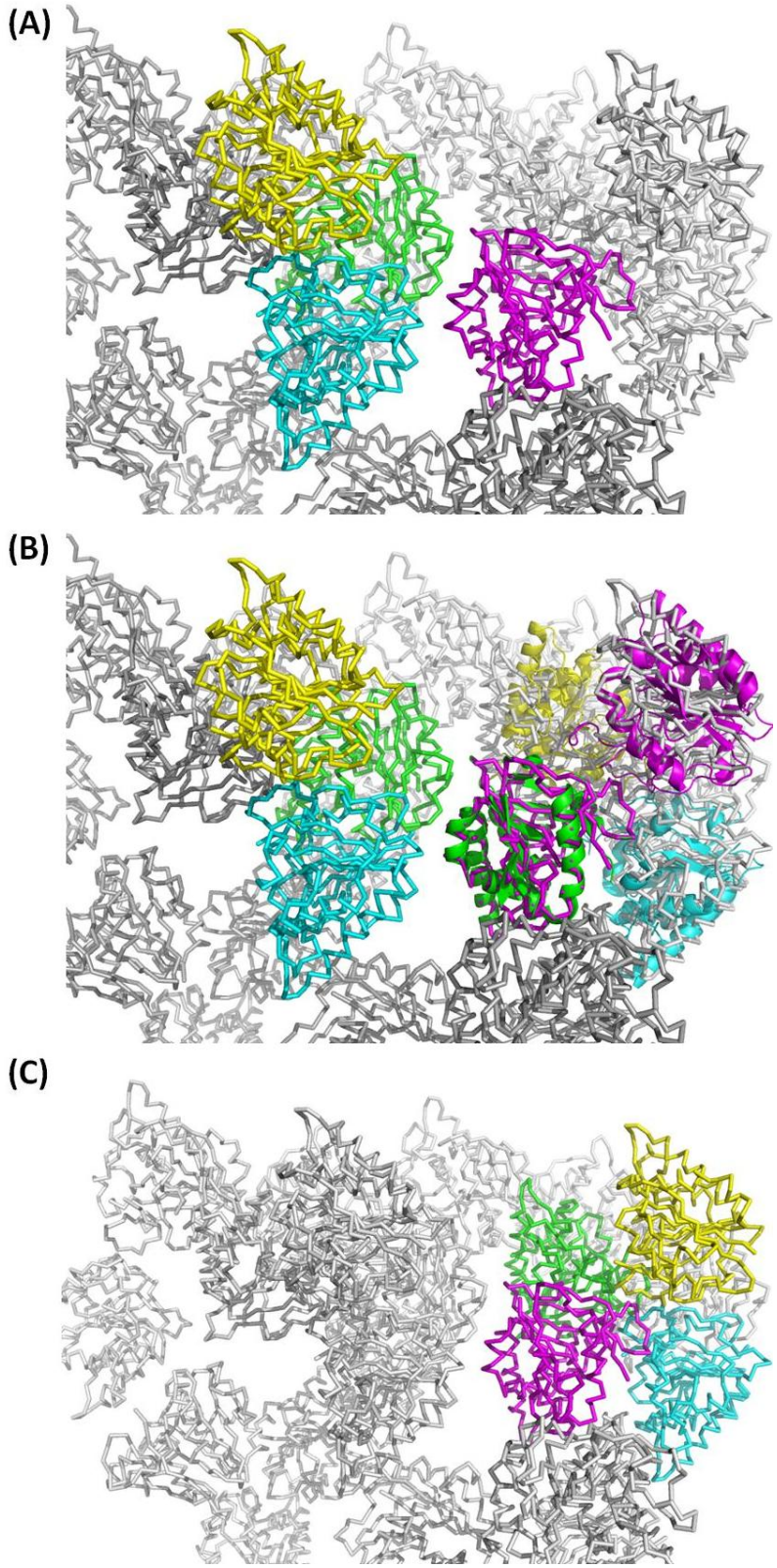


Figure 56. Comparison of the arrangement in the asymmetric unit of the BoGT6a E192Q•GalNAc structure and that of the BoGT6a•FAL structure. (A) 4 molecules of BoGT6a E192Q•GalNAc in the asymmetric unit with the presence of their symmetric

coordinates. The proteins are shown as ribbon and coloured by chain in which chain A is in green, chain B in cyan, chain C in yellow, and chain D in magenta. Their symmetric coordinates are also shown as ribbon, but coloured in grey. (B) a superposition of chain A of the BoGT6a•FAL structure with chain D of the BoGT6a E192Q•GalNAc structure. The BoGT6a•FAL structure is shown in cartoon representation in which molecules are coloured by chain. (C) new arrangement of 4 molecules in the asymmetric unit of the BoGT6a E192Q•GalNAc structure. Picture created using Pymol.

According to the relationships which were found amongst the form I structure, the form II structure and the acceptor bound structure, the same pseudo translation was also expected to happen with the molecules of the structure III. In fact, the comparison of the positions of molecules in the three structures showed that the arrangement of 16 molecules in the monoclinic form contained both packing styles from the other structures (Figure 57).

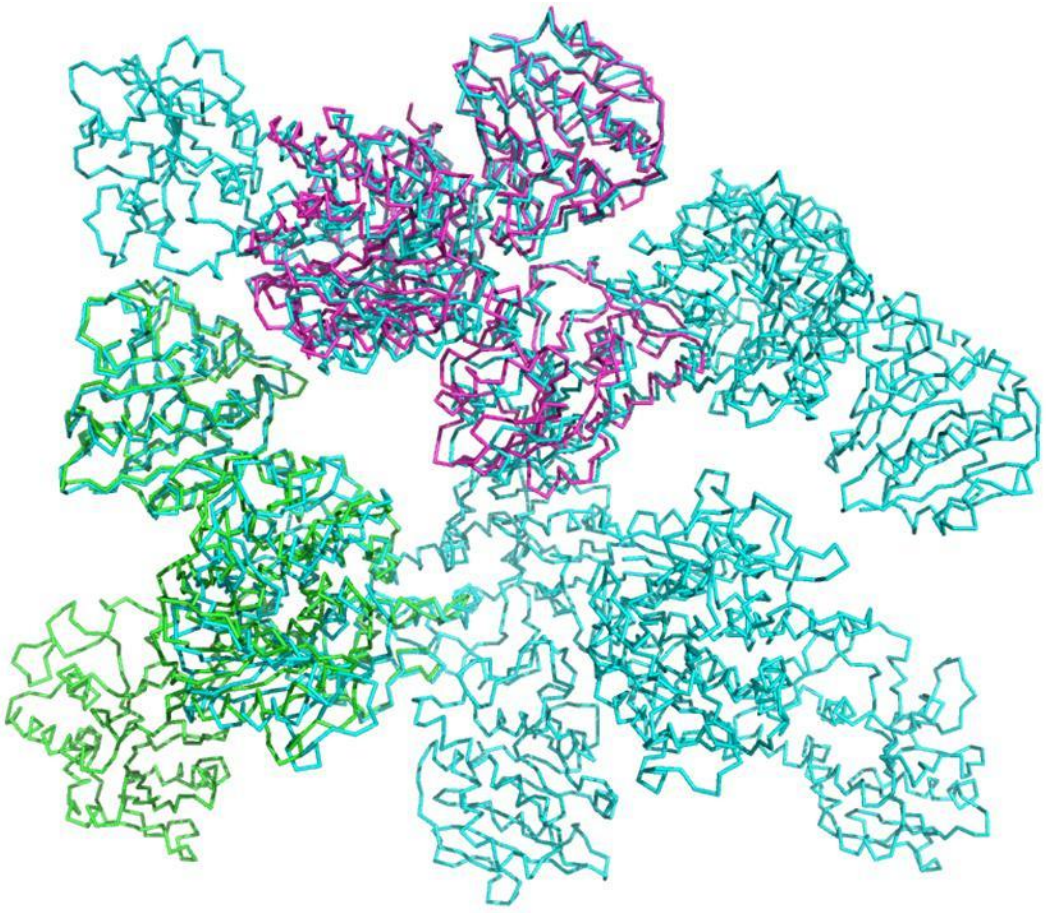


Figure 57. Comparison of the arrangement of molecules in the asymmetric unit of all three mutant BoGT6a E192Q complex structures. All the protein are shown as ribbon.

The BoGT6a•GalNAc structure is coloured in green, the BoGT6a•UDP-GalNAc form II structure in magenta, and the BoGT6a•UDP-GalNAc form III structure in cyan. Picture created using Pymol.

The same process of superpositions and analysis of the symmetry coordinates was performed with the BoGT6a E192Q•UDP-GalNAc structure in monoclinic form. It was superposed with the BoGT6a E192Q•UDP-GalNAc structure in the orthorhombic form because they had the same various phenomena of substrates bound in the active sites of its molecules. The symmetry coordinates of chains A, C, G, I, M and P were used instead of the molecules at their current positions, resulting in a long chain with 16 molecules composed of 4 groups, in which each group contained 4 molecules including 2 molecules with form B ligands, 1 molecule with form A ligand, and 1 molecule with form C ligand (Figure 58).

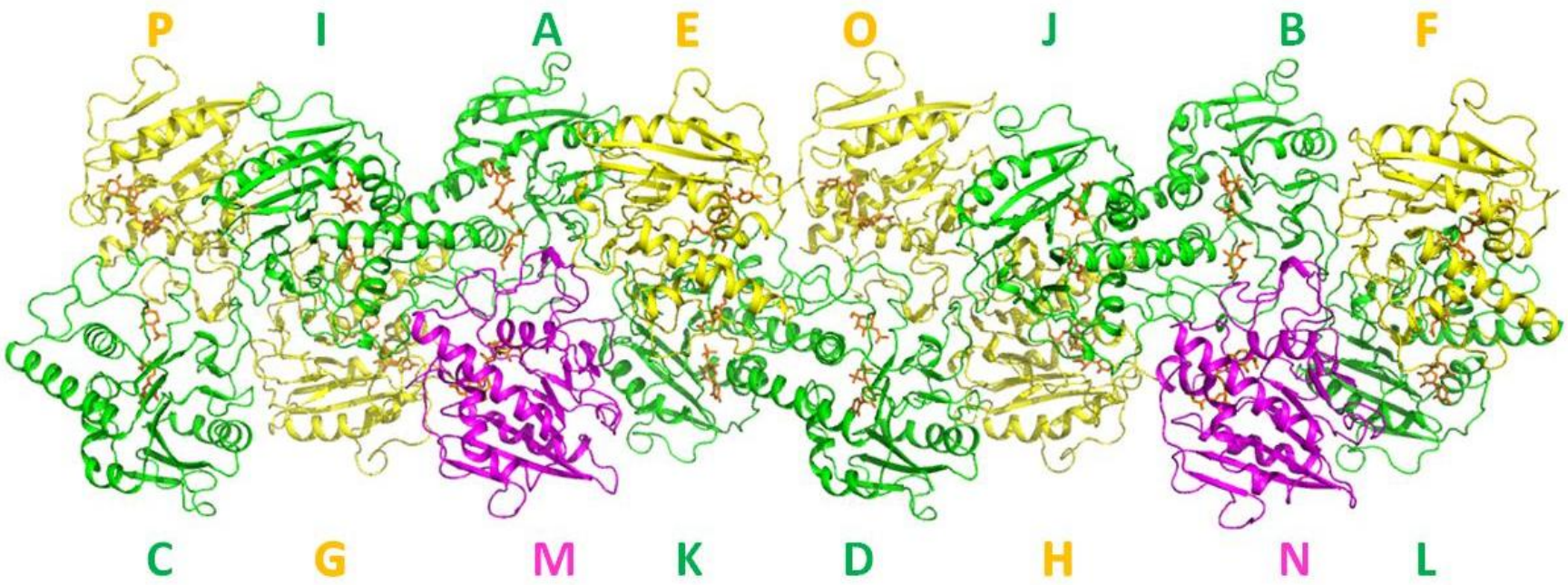


Figure 58. New arrangement of 16 molecules in asymmetric unit of the BoGT6a E192Q•GalNAc form III structure. The protein is shown as ribbon. Chains A, B, C, D, I, J, K and L which contain ligands in configuration B are coloured in green, chains E, F, G, H, O and P which contain ligands in configuration A in yellow, and chains M and N which contain ligands in configuration C in pink. Each chain is labelled with their name. Picture created using Pymol.

In all three crystal forms, the packing of molecules appears to be similar to that observed in the structure of BoGT6a in a complex with FAL. In both the form I structure and the BoGT6a•FAL structure there are two 2-fold axes of rotation. This is because the ligands in the active sites in each structure are similar. However, there is only one 2-fold axis in the form II structure due to the difference of the ligands in the active sites. The rotation exchanges molecules in pairs; one molecule in the pair has UDP-GalNAc and the other has UDP and GalNAc (Figure 59). Like in the BoGT6a in complex with FAL, the active sites of chain A and chain C face those of chain D and chain B respectively. The molecule with UDP and GalNAc is more solvent accessible with an exposed active site, while the active site of the molecule with UDP-GalNAc is buried by the C-terminal region.

Similar packing features are also observed in the monoclinic structure, but there is more elaborate packing which caused a high number of molecules in the asymmetric unit of this form because of the presence of the different structure (structure C). A self rotation function was performed using Molrep (Vagin and Teplyakov, 1997) for the form III structure with an integration radius of 3.61 Å. There are only peaks appearing at chi = 180^{o} and no significant peak at chi = 90^{o}, 120^{o} and 60^{o} (Figure 60). In the inspection chi = 180^{o}, two strong peaks on the y axis at the origin indicate the crystallographic 2-fold axis along the z axis. There are two extra peaks on the x axis which are at 42.19^{o} apart from the z axis (Figure 60). This indicates a NCS with a 2 fold axis which is parallel to the (x, z) plane and 42.19^{o} different from the z axis of the unit cell (Figure 61).

As superposition shows, the core group of the form III structure is 4 molecules in which a molecule with the ligand in structure A is grouped with a molecule with the ligand in structure B. There is, however, a new group of 4 molecules in which 1 pair is composed of a molecule with ligand in structure A and a molecule with ligand in structure B, and the other pair is composed of a molecule with ligand in structure C and a molecule with ligand in structure B. A pseudo translation along the x axis causes the number of molecules in the asymmetric unit to double from 8 to 16 molecules (Figure 61). Among the 4 molecules of each group, molecules in a pair are mediated by contacts between residues Glu223-Lys233, Pro221-Gly227 and each pair interacts with the other pair through Asn127 and Leu219 (Figure 62).

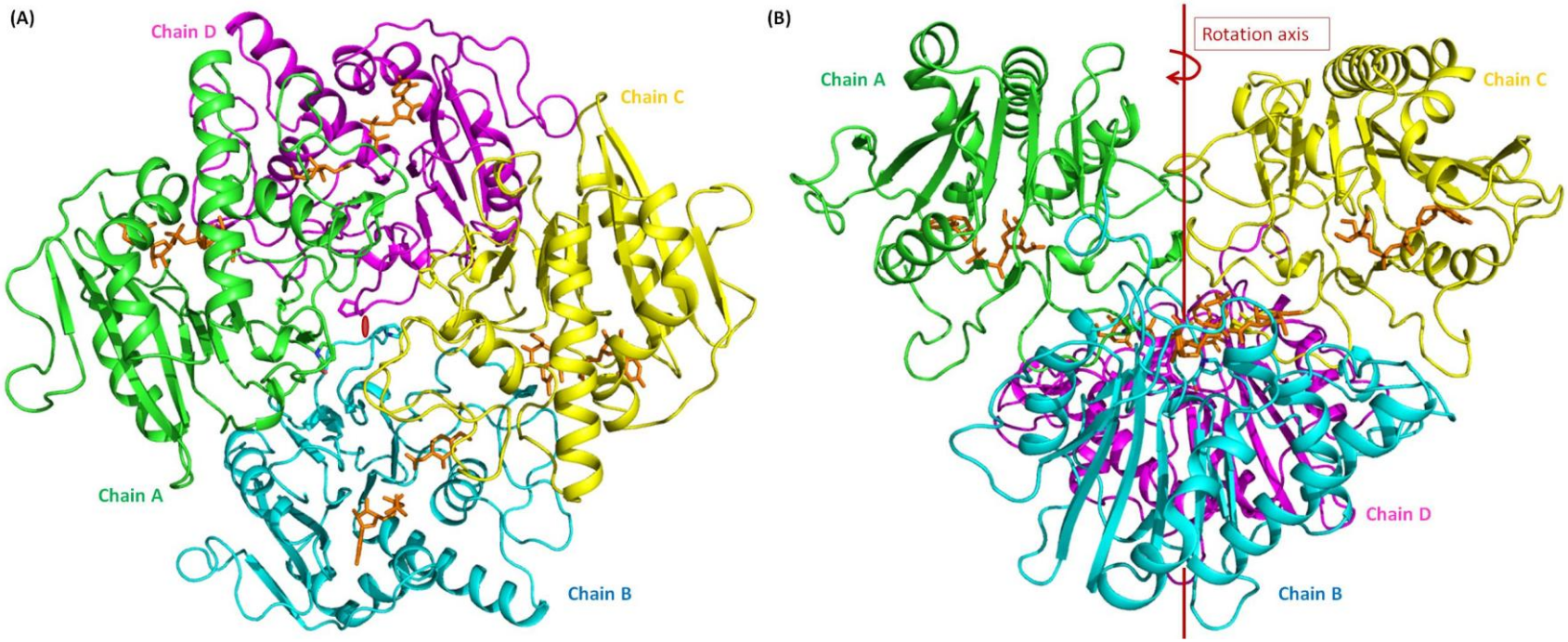


Figure 59. Symmetry in the BoGT6a E192Q•UDP-GalNAc complex form II structure. (A) Top view shows the rotation operation applied to the structure. The 2-fold symmetry axis was marked by a red oval in the center. (B) shows the side view of the structure and the rotation axis. The protein is shown in cartoon representation where chain A is coloured in green, chain B in cyan, chain C in yellow and chain D in magenta. The residues Pro221 are shown as stick and coloured following the chain that they belong to. The ligands are shown as orange sticks. Picture created using Pymol.

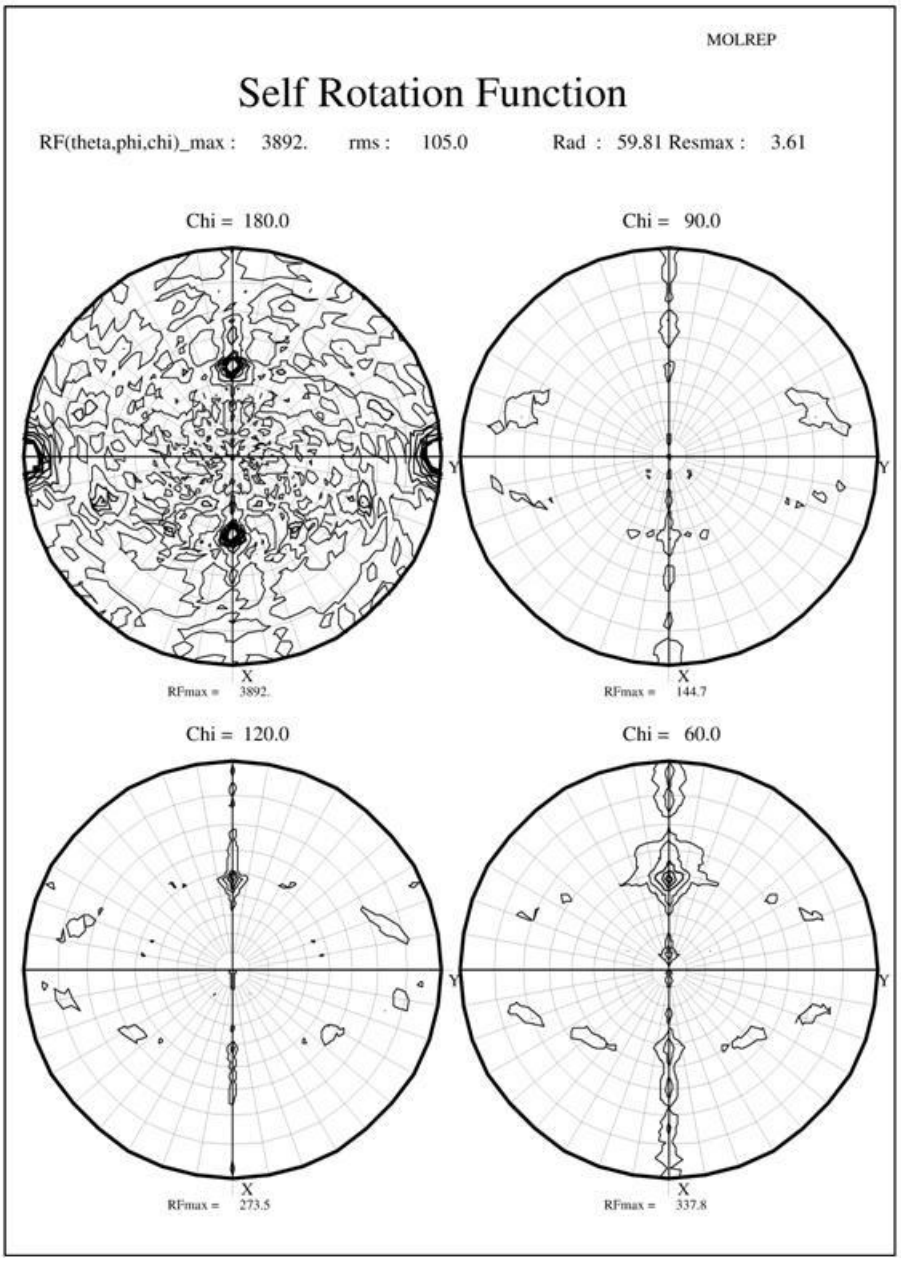


Figure 60. Self rotation function result for the BoGT6a E192Q•UDP-GalNAc form III structure. The chi angles are shown. There are two peaks at the origin clearly indicating the crystallographic 2-fold axis and two extra peaks around 42^{o} apart from the y axis indicating the non-crystallographic 2-fold axis. Picture created using Molrep (Vagin and Teplyakov, 1997).

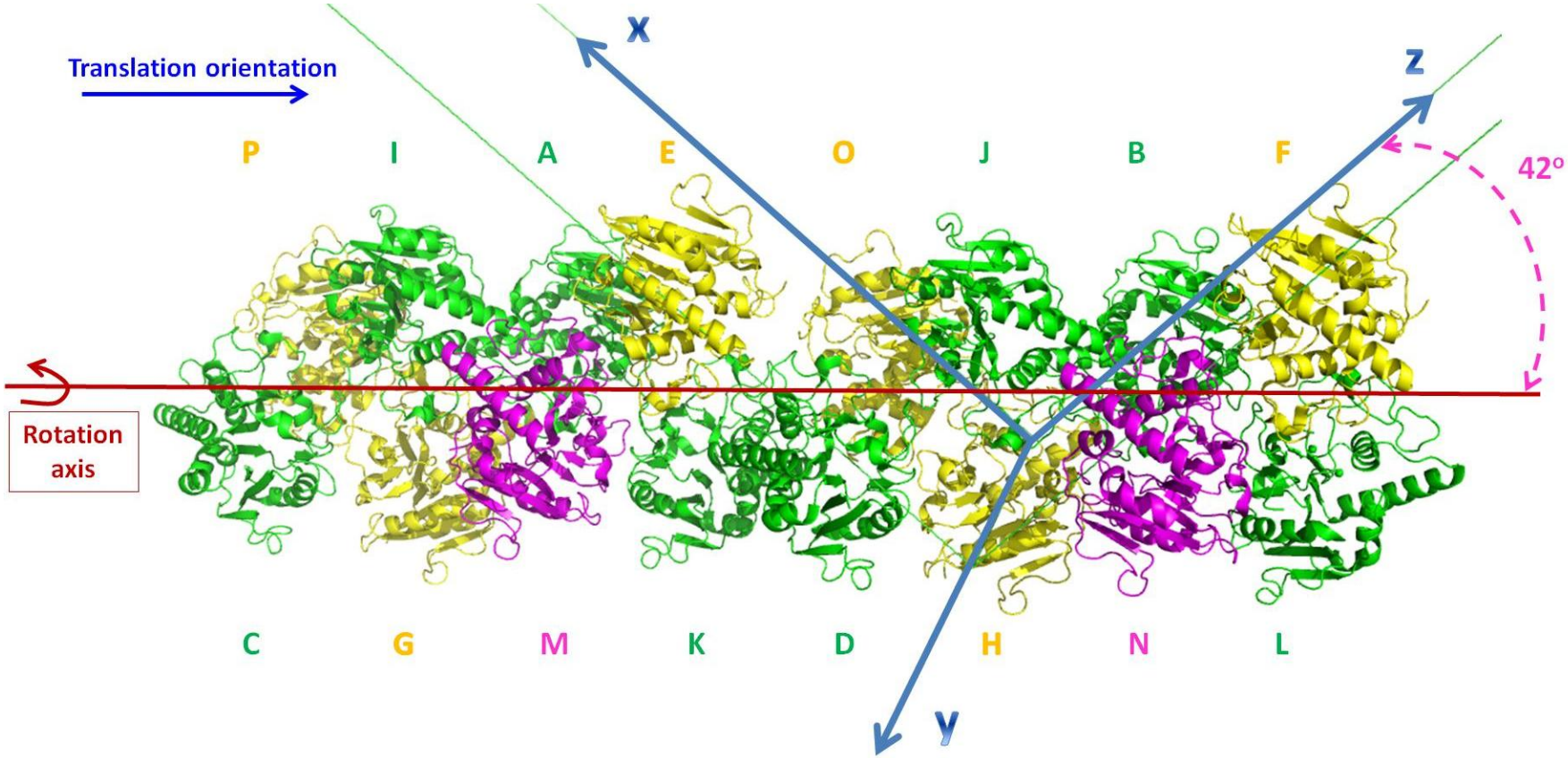


Figure 61. Symmetry in the BoGT6a E192Q•UDP-GalNAc form III structure. The protein is shown in cartoon representation where molecules with ligands in configuration A are coloured in yellow, configuration B in green, and configuration C in magenta. The unit cell is shown in green. The x, y and z axes are shown as blue arrows and noted. The translation direction and the rotation axis are noted. The deviation angle between the z axis and the rotation axis of the NCS is noted. The picture created using Pymol.

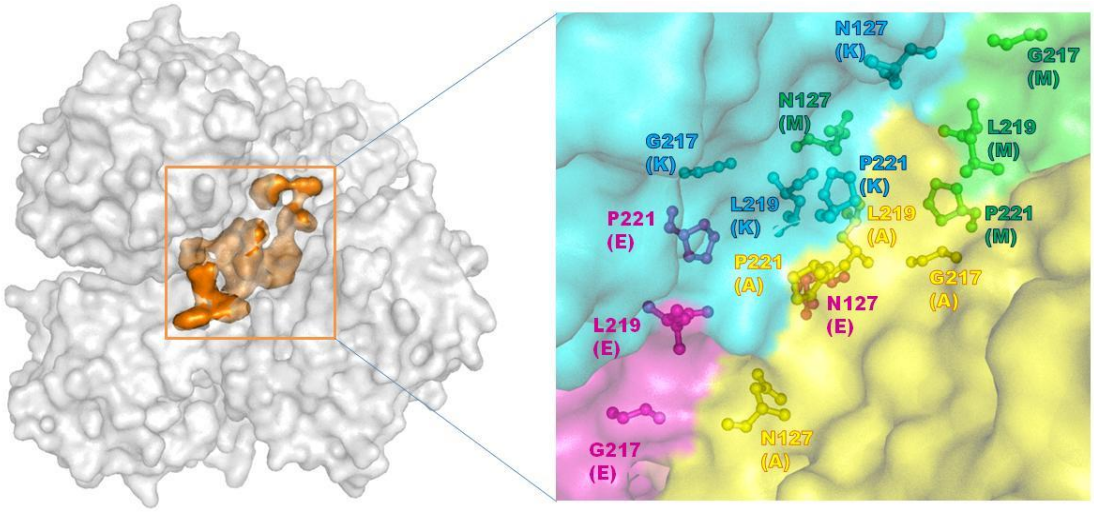


Figure 62. Arrangement and interactions among chains in the form III structure. The protein is showed as surface in grey and the hydrophobic core in orange. The inset shows a closer look in hydrophobic core in which hydrophobic residues are shown as stick and colored by chains. Picture created using Pymol.

3.3.3.2 Interactions with ligands and conformational changes

Like the acceptor bound structure, the overall structures of the complexes are similar to that of the apo protein, which follow the GT-A fold comprising two contiguous subdomains. The first is the N-subdomain (residues 1-94) that includes 4 parallel β-strands and 3 surrounding α-helices. The other part is the C-subdomain (residues 95-246) that includes a 3 stranded mixed β-sheet and a small 2 stranded β-sheet associated with two α-helices.

The difference between these structures compared to the apo form and the acceptor bound form is the length of the C-terminal region, which is longer due to the increased order of this region when the enzyme interacts with the donor substrate (Figure 63). The restructuring of the C-terminus of the enzyme was also observed in the BoGT6a in complex with the acceptor substrate, which is considered as a "closed" form of the enzyme. In this form, the binding sites of both UDP-GalNAc and FAL are located in a pocket on the protein surface that is covered by "a lid" formed by the C-terminus. This form is in contrast to the "open" form of the BoGT6a apo structure in which the C-terminus is flexible and in an orientation pointing out of the active site (Figure 63).

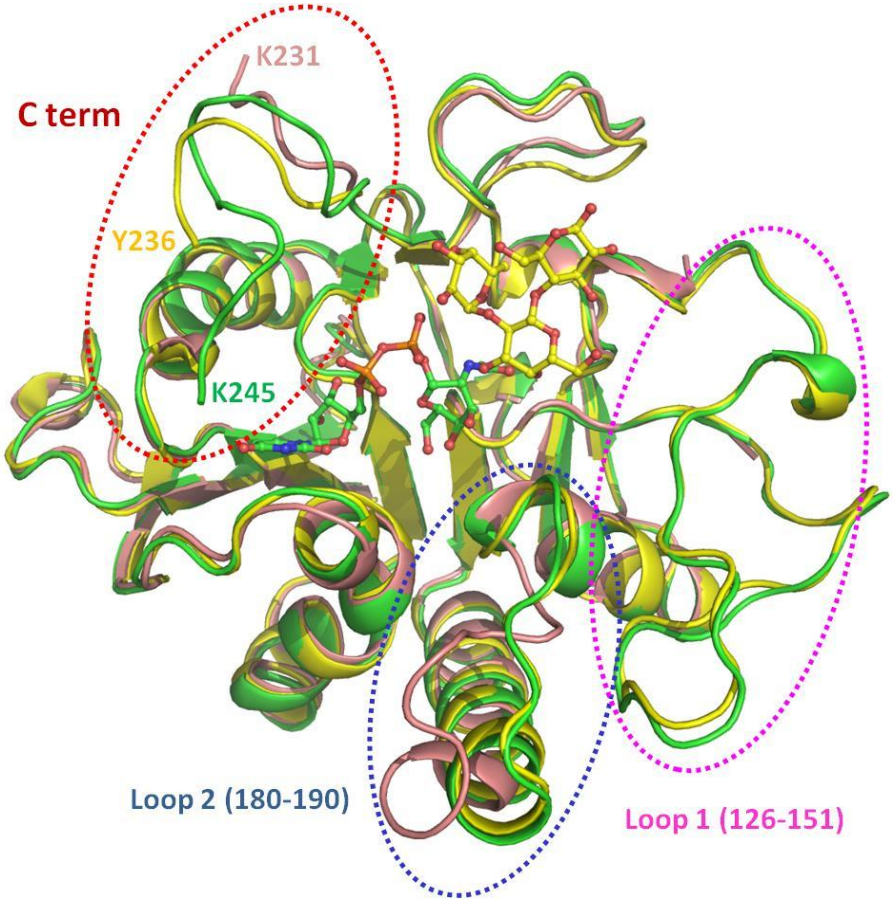


Figure 63. Conformational comparison of the overall structures of the BoGT6a apo form structure (in pink), the BoGT6a•FAL structure (in yellow) and the BoGT6a E192Q•UDP-GalNAc form III structure (in green). The proteins are shown in cartoon representation. UDP-GalNAc is shown as green sticks and FAL as yellow sticks. The conformational differences are marked and labelled. The end residue of each chain is noted in colour as its protein. Picture created using Pymol.

The interaction with either substrate stabilises loop 1 (residues 126-150) that is unstructured in the apo form and induces the enzyme to undergo a conformational change to a less open form in which loop 2 (residues 180 – 192) also changes conformation compared to the apo form structure (Figure 63).

Although the acceptor bound structure has a closed structure, the C-termininus beyond residue Lys231 is slightly flexible with only one of 4 molecules in the asymmetric unit showing clear electron density up to residue Tyr236. In the form I structure, this C-terminal region is more stable but as in the BoGT6a•FAL structure, the structure could not be traced beyond residue Tyr236. Superposing these two

structures showed that the position of the GalNAc moiety is the same as that of the FAL moiety in the acceptor binding site, but there are fewer interactions between the GalNAc and the enzyme than between the FAL and the enzyme (Figure 64). A close look at the interactions between the enzyme and GalNAc shows that no residue beyond Lys231 interacts with the sugar moiety. This is in agreement with the idea discussed above that the C-terminal region is not involved in acceptor interaction, leading to its high flexibility when there is only GalNAc moiety remaining in the active site.

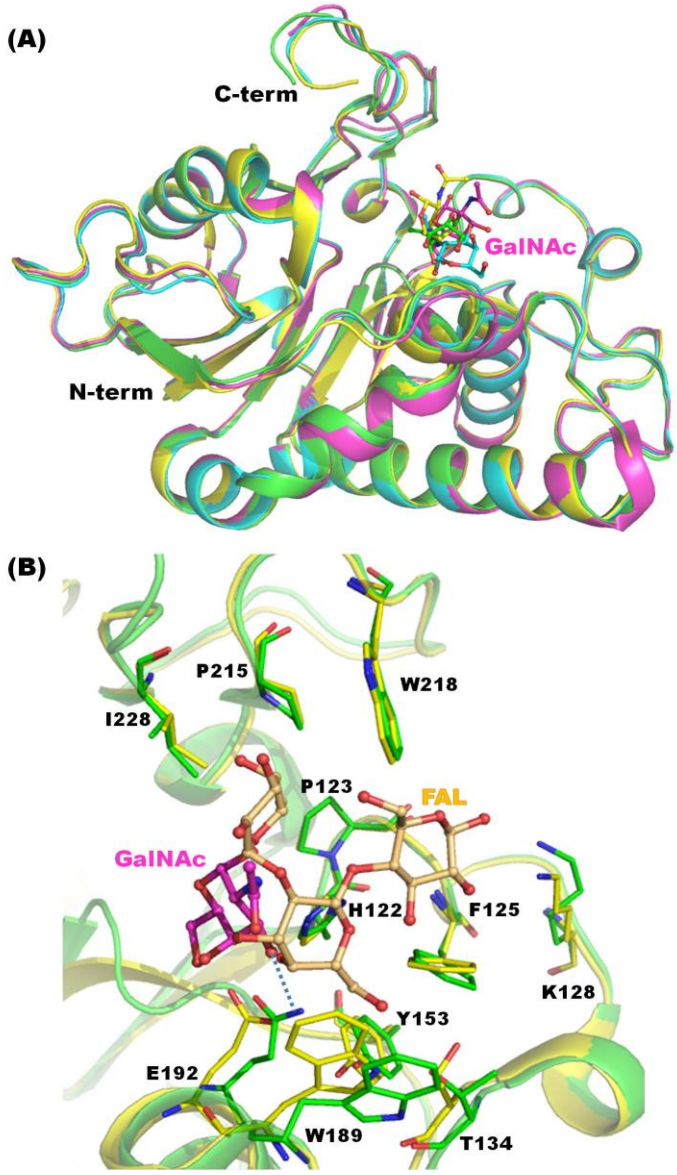


Figure 64. Comparison of the active sites of the BoGT6a•FAL structure and the BoGT6a•GalNAc structure. (A) a superposition of 4 molecules in the asymmetry of the

BoGT6a•GalNAc structure. The proteins are shown in cartoon representation where chain A is coloured in green, chain B in cyan, chain C in yellow and chain D in magenta. α-GalNAc moieties (noted as GalNAc) are shown as sticks and coloured according to the molecule they belong to. C-terminus (noted as C-term) and N-terminus are marked. (B) a superposition of the active sites of the BoGT6a•FAL structure and the BoGT6a•GalNAc structure shows GalNAc in the acceptor binding site. The proteins are shown in cartoon representation and coloured in yellow for the BoGT6a•FAL structure, and in green for the BoGT6a•GalNAc structure. FAL is shown as light orange sticks and GalNAc as magenta sticks. Residues in the acceptor binding site are shown as lines and coloured according to the structure they belong to. H-bond is shown as a blue dash line. Pictures created using Pymol.

In the form II structure and the form III structure, interactions with the donor substrate, especially the interaction between Lys231 and the diphosphate moiety of the UDP-GalNAc or UDP, stabilise part of the C-terminus and the structures can be followed beyond residue Lys236. However there are some variations of the C-terminal regions among different chains beyond this residue which are related to the different configurations of their ligands, suggesting that this region is flexible (Figure 65).

The C-terminal region could be followed up to residue Leu241 (for the form II structure) or Lys245 (for the form III structure) in the molecules where the UDP-GalNAc is intact or the sugar moiety has not been moved to the acceptor binding site (Figure 65). As in the form I structure, the C-terminus becomes less ordered when UDP-GalNAc has been hydrolysed and GalNAc has moved to the acceptor binding site, leading to an absence of electron density for the C-terminal 10 residues and hence the models were built only up to residue Tyr236. This indicates that the C terminal region mainly involves in donor binding activity of the enzyme.

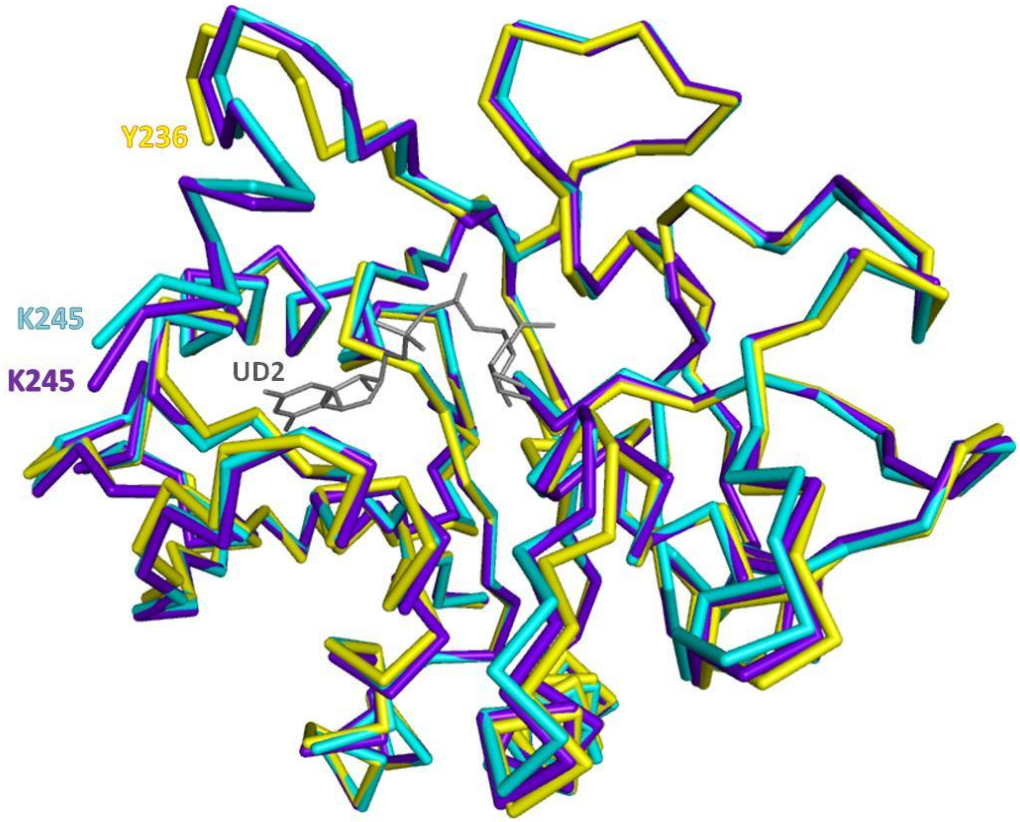


Figure 65. A superposition of three configurations of the BoGT6a E192Q•UDP-GalNAc form III structure. The protein is shown in ribbon representation where the configuration A (chain E as representative) is coloured in cyan, configuration B (chain A as representative) in yellow, and configuration C (chain M as representative) in purple. The UDP-GalNAc ligand from chain E is coloured in grey to show the relation between the C terminus and the donor substrate. The end residue of each chain is labelled. Picture created using Pymol.

The surface diagrams of the three forms in the monoclinic structure show that C (chain M) is slightly more open than A (chain E) while B (chain A) is the most open form in which both the UDP and GalNAc ligands are exposed to solvent (Figure 66).

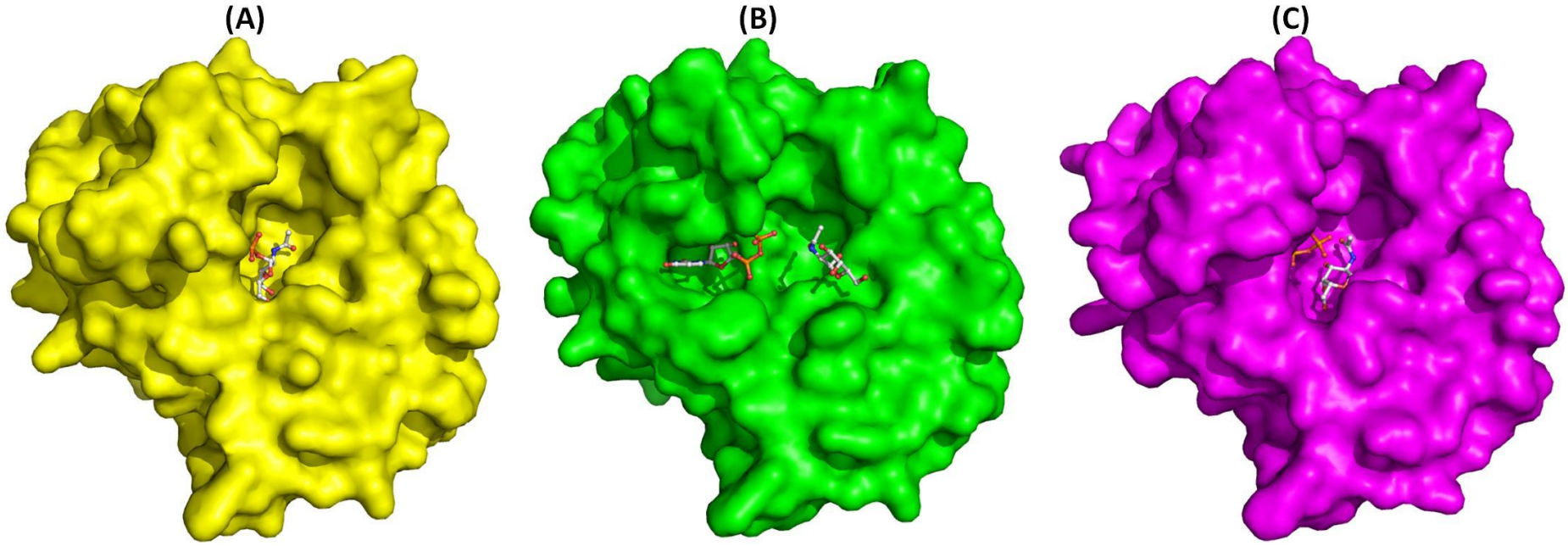
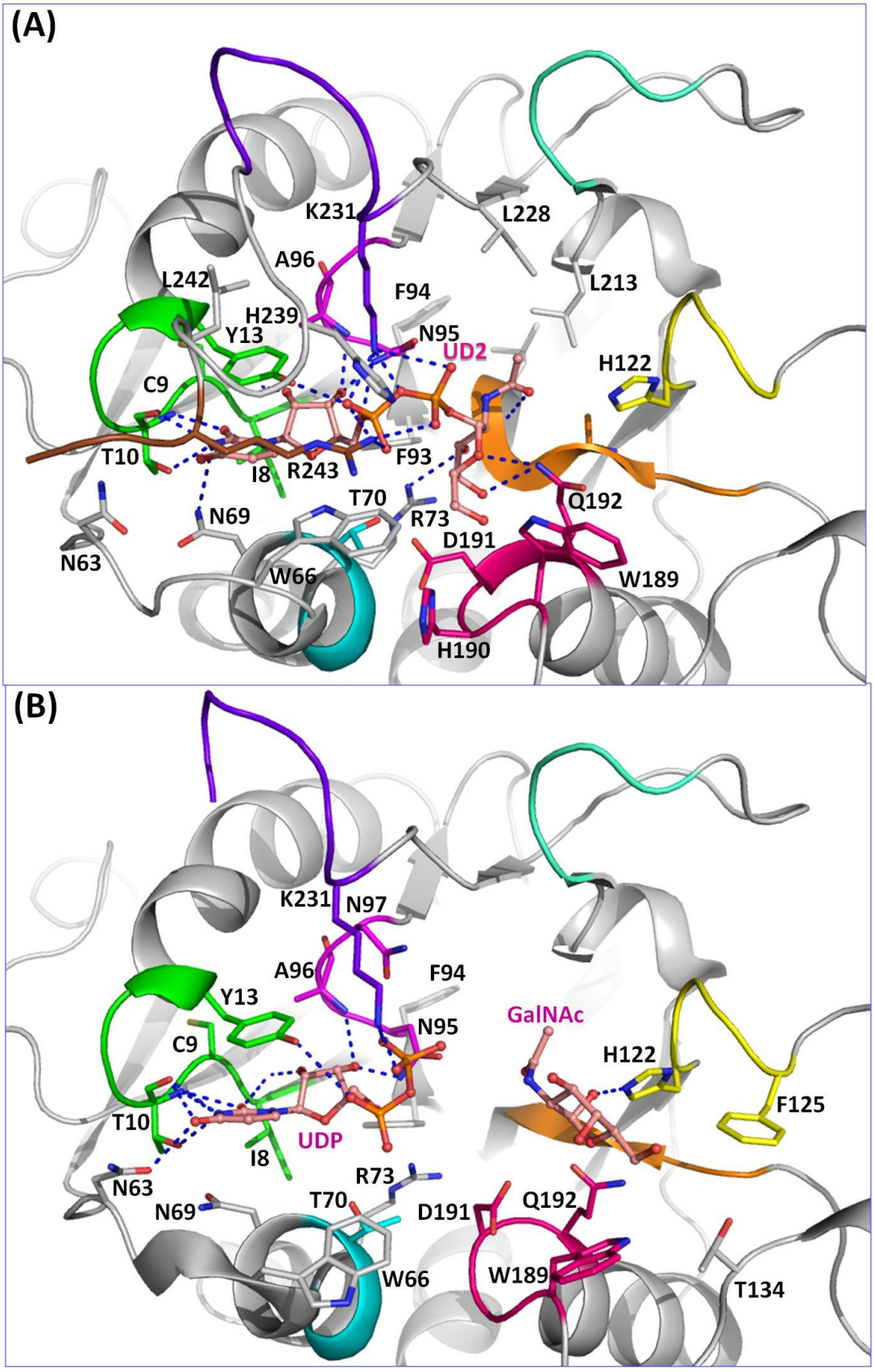


Figure 66. Surface diagrams of three BoGT6a E192Q donor bound structures in monoclinic form. (A) configuration A (yellow), (B) configuration B (green), and (C) configuration C (magenta). The ligands are shown as grey sticks. Picture created using Pymol.

Insight into the interactions between the enzyme and its donor substrates in the configuration A and configuration C of the form II structure shows that residues Lys231, His239 and Arg243 belonging to the C terminus directly interact with the UDP moiety of the donor substrate, UDP-GalNAc in the configuration A or the UDP in the configuration C (Figure 67A and C). These interactions induce the conformational change of the C terminal region and also stabilise it. This explains why the C terminus can be built almost completely (only Asn246 was not built) in these structures, but incompletely in the acceptor bound structure, the form I structure and in the configuration B of the form III structure.

In the configuration A of the form III structure, the ε-amino group of Lys231 H-bonds with O1A, O1B and O3A atoms and the NH_2 of Arg243 interacts weakly with different oxygen atoms in different chains. Furthermore, the OH of Tyr13, and peptide N of Ala96 make H-bond interactions with the diphosphate (Figure 67A, Figure 68A). On the other hand, in the configuration C of the form III structure, the amino group of Lys231 interacts more strongly with O1A and also with O1B of the diphosphate and the NH_2 of Arg243 is near to but not in H-bonding distance (3.66 Å) of O2A. Ala96 interacts with O3B but there is a stacking interaction between Tyr13 and the substrate (Figure 67C, Figure 68C). The β-GalNAc C1 is in close contact with Gln192 NE2 (Figure 67C, Figure 68C).

In the configuration B of the form III structure, whilst there are similar interactions between the protein and the uracil and ribose, the diphosphate has a variable orientation in different chains and the Lys231 NH_2 and Tyr13 OH interact with different oxygens, although the Ala96 NH and Asn95 ND2 mainly interact with O3B. The GalNAc moiety interacts with Gln192 NE2 through O4' and Arg73 NH_2 and Gly157 N through O3' as well as Asn95 OD1 through N2' (Figure 67B, Figure 68B). Phe125 and Thr134, which are conserved in the acceptor binding sites of GT6 family members, appear as interacting residues with the GalNAc moiety. These residues do not involve in the donor substrate binding stage (the configuration A and C). This indicates that GalNAc moiety was transferred from the donor binding site to the acceptor binding site. In other words, the configuration B of the form III structure contains the product of the hydrolysis of UDP-GalNAc.


(A)
K231
L228
A96
L242
F94
L213
H239
Y13
N95
UD2
H122
C9
T10
I8
R243
F93
T70
Q192
N69
R73
D191
N63
W66
W189
H190
(B)
K231
N97
A96
F94
Y13
GalNAc
N95
C9
H122
F125
T10
UDP
I8
R73
N63
T70
D191
Q192
N69
W66
W189
T134

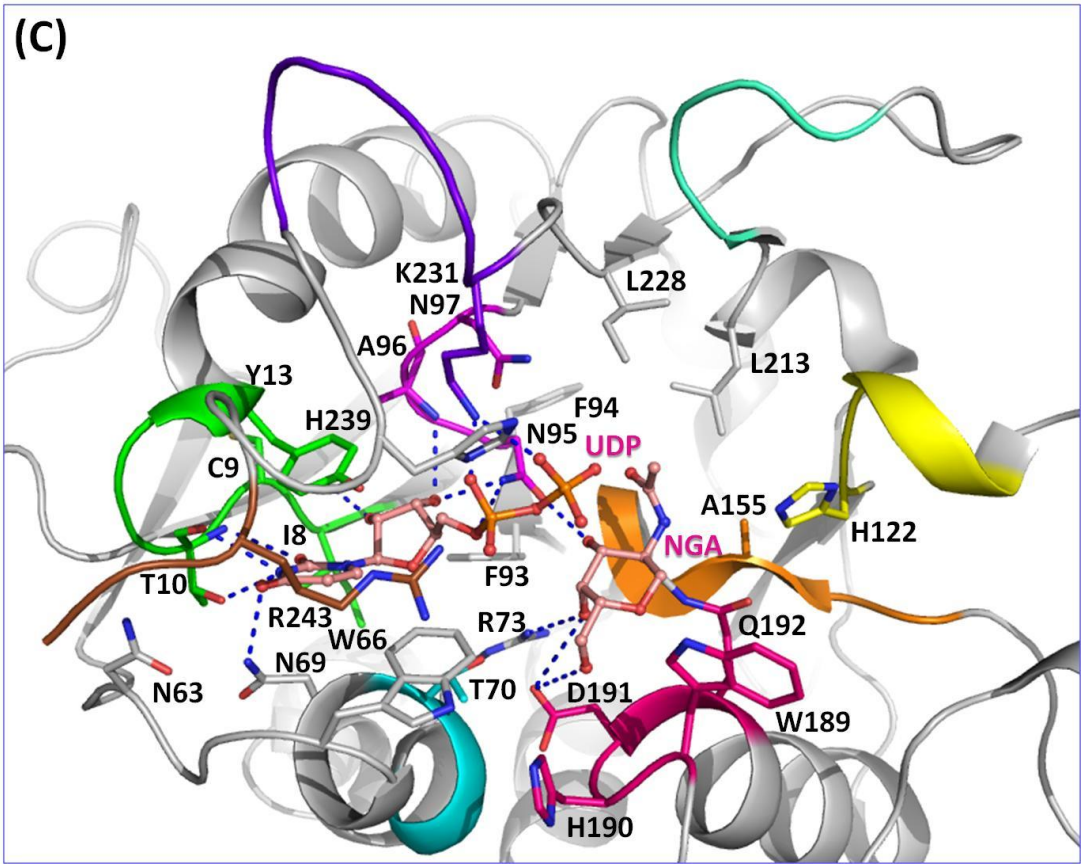


Figure 67. Interactions of BoGT6a E192Q with bound ligands in the form III structure. (A) with ligand in the configuration A, (B) with ligand in in the configuration B, and (C) with ligand in the configuration C. The ligands are shown as pink sticks and marked as UD2 for UDP-GalNAc, UDP for UDP, GalNAc for α-GalNAc and NGA for β-GalNAc. The interacting residues are shown as line and labelled in 1 letter abbreviation. The protein is shown in a cartoon representation where LBR-A is coloured in green, LBR-B in cyan, LBR-C in magenta, LBR-D in yellow, LBR-E in orange, LBR-F in hot pink, LBR-G in green cyan, LBR-H in purple blue, and LBR-I in brown. Picture created using Pymol.

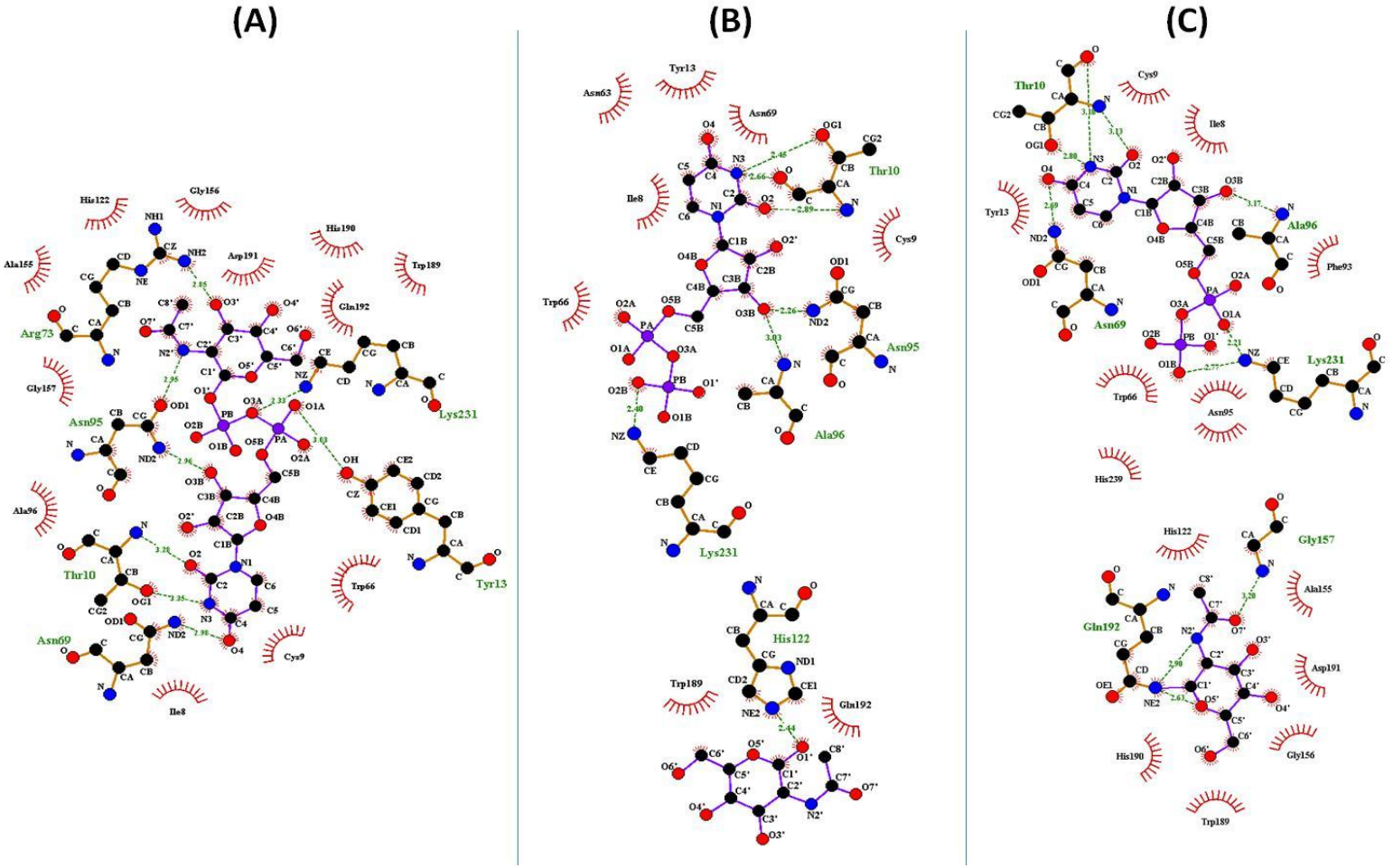
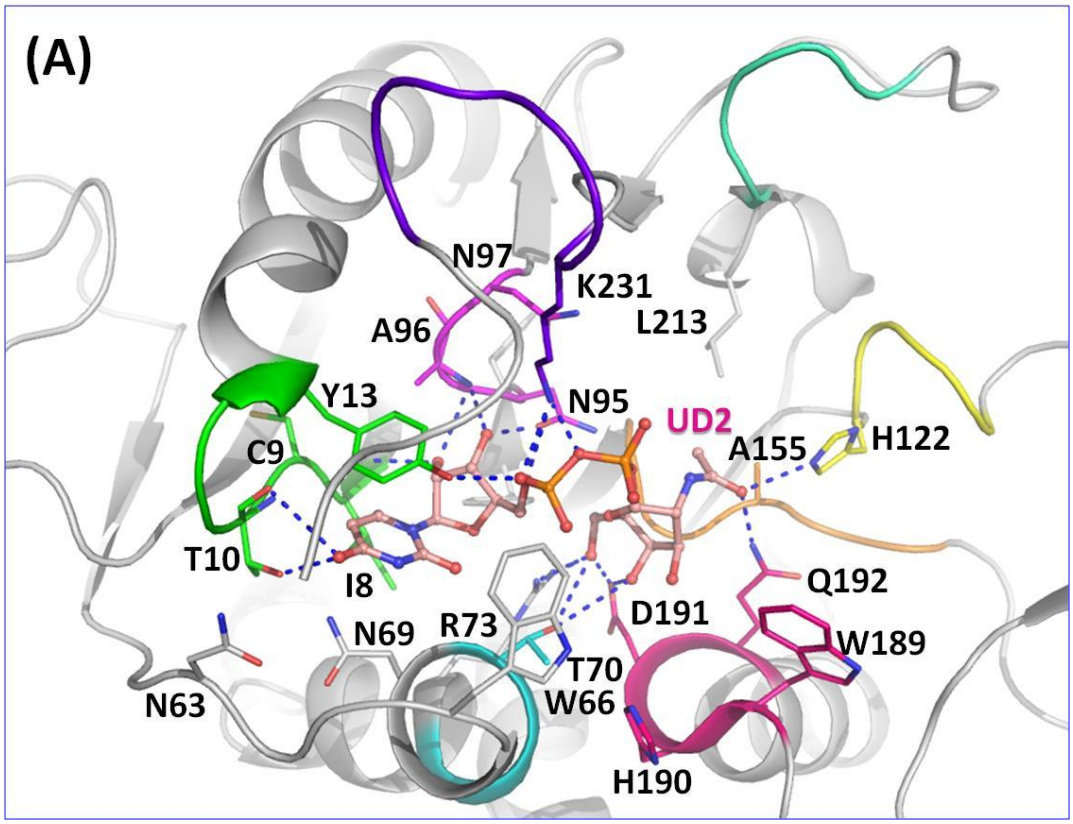


Figure 68. Ligplot of BoGT6a E192Q-ligand interactions with key residues in different complexes. (A) structure A, (B) structure B and (C) structure C. The ligands are shown in purple, interacting residues in orange and hydrophobic interacting residues in red colours. Hydrogen bonds are shown as green dashes. Image created using Ligplot (Wallace *et al.*, 1995).

A comparison of the interactions between the enzyme and its ligands in the form II and those in the form III structures shows that the interactions are similar between the two forms, apart from the lack of interactions between Arg243 and the phosphate group of the UDP-GalNAc in the form II structure (Figure 69). This is due to the lower quality of this structure, leading to insufficient electron density for residues beyond Leu241. As mentioned above, only the configuration A and B are observed in the form II structure. The configuration C in the form III structure is assumed to be an intermediate stage of the enzyme catalytic process; it exists only for a brief time, and so is difficult to detect. Hence, for further discussion, the form III structure is used as representative of the structure of BoGT6a E192Q in complex with the donor substrate. This is because it contains the most information about the catalytic process of the enzyme.



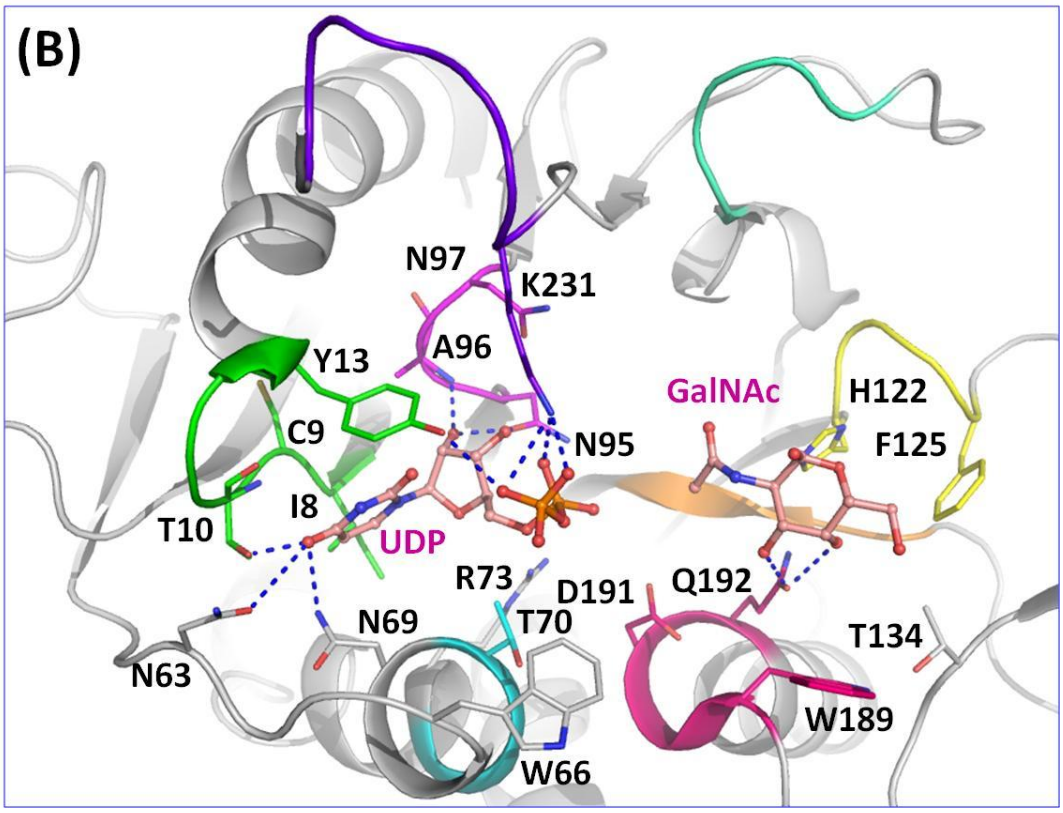


Figure 69. Interactions of the BoGT6a E192Q with its ligands in the form II structure. (A) with ligand in the configuration A, and (B) with ligand in the configuration B. The ligands are shown as pink sticks and marked as UD2 for UDP-GalNAc, UDP for UDP and GalNAc for α-GalNAc. The interacting residues are shown as line and labelled in 1 letter abbreviation. The protein is shown in a cartoon representation where LBR-A is coloured in green, LBR-B in cyan, LBR-C in magenta, LBR-D in yellow, LBR-E in orange, LBR-F in hot pink, LBR-G in green cyan, LBR-H in purple blue, and LBR-I in brown. Picture created using Pymol.

3.3.3.3 Proposed mechanism of the hydrolysis by BoGT6a

Although the structure of a ternary complex of BoGT6a with both UDP-GalNAc and FAL has not been determined, an examination of the structures with individual substrates still provides some information about the enzyme catalytic mechanism.

In the structure of BoGT6a apo form structure, the C terminus is orientated in a direction away from the active site of the enzyme (Figure 70A). This makes the active site exposed completely to the environment, known as an open conformation

of the enzyme. Such an open state is available to accommodate both the acceptor substrate and the donor substrate (Figure 70B).

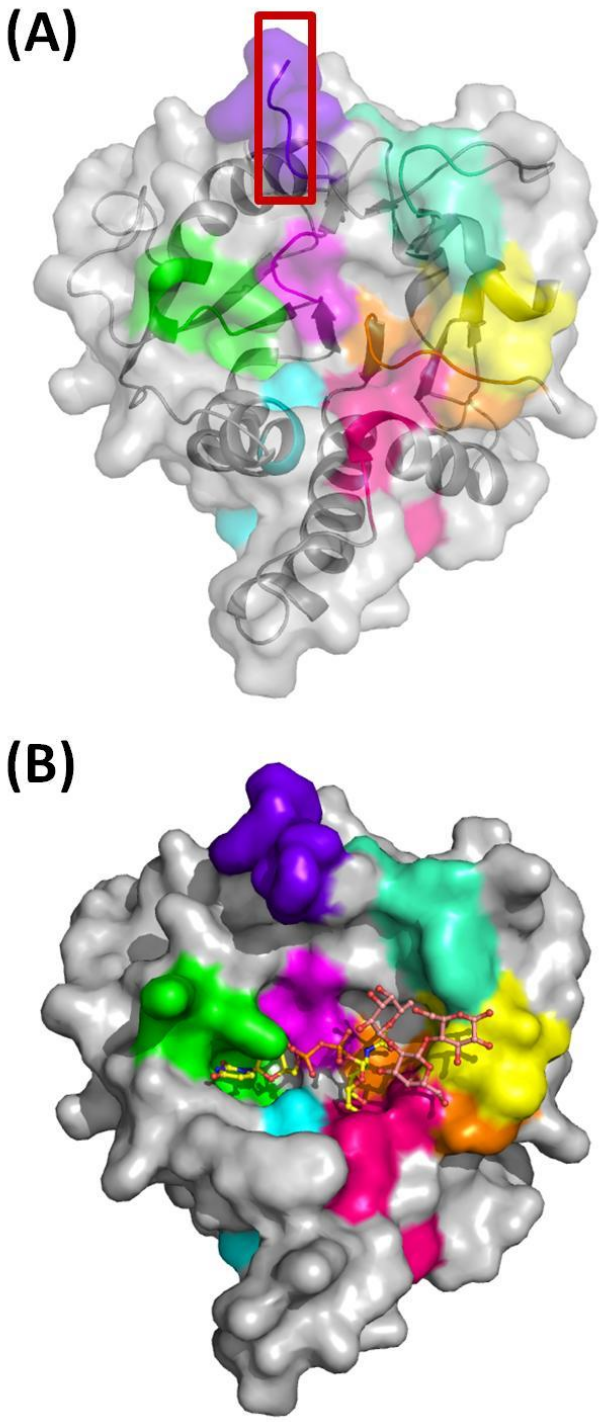
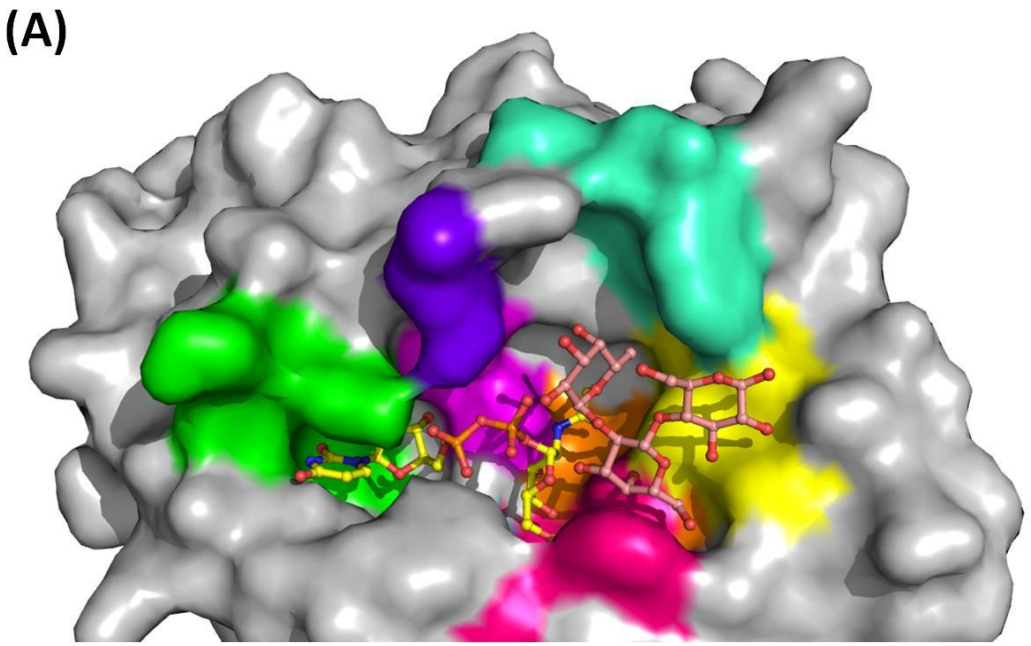


Figure 70. Analysis of surfaces of the BoGT6a apo form. (A) the surface of the BoGT6a apo form (PDB 4AYL) shows an open active site with the flexible C-terminus as an open lid. The end of the C terminus is boxed. The protein is shown in cartoon representation. (B)

the surface of the model of BoGT6a apo form with both FAL and UDP-GalNAc in the configuration A shows an open active site available for all the ligands. LBR-A is coloured in green, LBR-B in cyan, LBR-C in magenta, LBR-D in yellow, LBR-E in orange, LBR-F in hot pink, LBR-G in green cyan, and LBR-H in purple blue. Picture created using Pymol.

However, the access way of the active site is restricted when the enzyme interacts with its ligands. A model of the structure of BoGT6a in complex with its acceptor substrate, FAL and the UDP-GalNAc shows that the active site is not accessible for the donor substrate UDP-GalNAc. However, the FAL binding site is accessible for FAL in the complex with UDP-GalNAc in the model of BoGT6a E192Q in complex with UDP-GalNAc (configuration A) and FAL. Previous ITC studies that show FAL binds weakly to free BoGT6a (K_d 1.2 mM) but more strongly to the UDP complex (K_d 76 μM); the change in free energy of binding (-1.64 kcal/mol) arises from a more favorable enthalpy of binding (δΔH of -1.9 kcal/mol) (Thiyagarajan *et al.*, 2012). This finding suggests BoGT6a follows a Bi-Bi sequential kinetic mechanism in which the donor substrate binds to the enzyme before the acceptor substrate, which is found in many glycosyltransferase (Rini *et al.*, 2009).



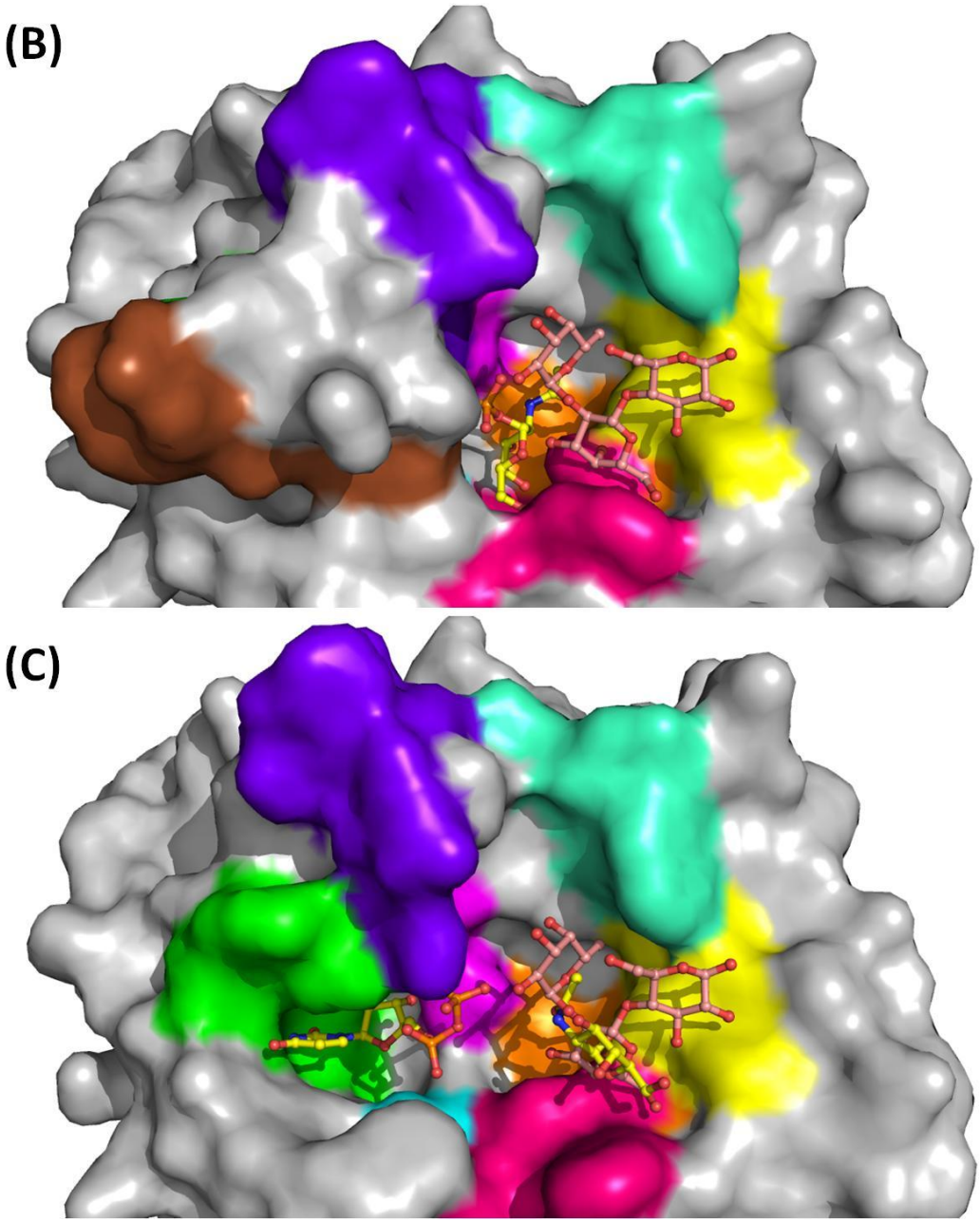


Figure 71. Analysis of surfaces of the BoGT6a complex forms. (A) shows the surface of the BoGT6a•FAL modelled with UDP-GalNAc in the configuration A. The access of the donor binding site is restricted. (B) shows the surface of the BoGT6a E192Q•UDP-GalNAc in the configuration A modelled with FAL. The acceptor binding site is still accessible. (C) shows the surface of the BoGT6a E192Q•UDP-GalNAc in the configuration B modelled with FAL. The active site is accessible for both the donor substrate and the acceptor substrate. UDP-GalNAc, UDP and α-GalNAc are shown as yellow sticks and FAL as pink sticks. LBR-A is coloured in green, LBR-B in cyan, LBR-C in magenta, LBR-D in yellow,

LBR-E in orange, LBR-F in hot pink, LBR-G in green cyan, LBR-H in purple blue and LBR-I in brown. Picture created using Pymol.

In addition, the presence of three configurations of the ligands in the form III also provides a nice picture of the hydrolysis reaction catalysed by BoGT6a.

The configuration A of the form III structure represents the UDP-GalNAc Michaelis complex with BoGT6a, but it should be noted that the E192Q mutation in the form of BoGT6a used in these structural studies involves a key residue in catalysis (Tumbale and Brew, 2009), that interacts with both the donor and acceptor substrates. Previous ITC studies have shown that this mutation has little net effect on the ΔG for UDP-GalNAc binding (both around -5.8 kcal/mol) but this reflects mutually compensating effects on the ΔH and TΔS of binding (changes from -23.5 to -16.4 kcal/mol and 17.7 to 10.6 kcal/mol, respectively) suggesting that the mutation may weaken non-covalent interactions and reduce substrate-induced conformational rearrangements (Thiyagarajan *et al.*, 2012). This agrees with the lower catalytic rate of the mutant compared to that of the wild type. This configuration shows the first step of the hydrolysis reaction when the intact UDP-GalNAc binds to the enzyme.

The configuration B of the BoGT6a E192Q form III structure shows an open active site which is accessible for both the acceptor and donor substrate, or in other words, is ready to release the product. This can be the last step when the hydrolysis reaction is complete.

The presence of the configuration C structure, which presents a glycosyl enzyme with a covalent bond between the residue Glu192 and C1 of the GalNAc moiety of the UDP-GalNAc, can be considered as the intermediate stage of the enzyme catalytic process. This finding is structural evidence supporting the double displacement mechanism for BoGT6a catalytic activity. Following this mechanism the mutant BoGT6a E192Q still retains its activity because amine group of Gln can play a role as a catalytic nucleophile as hydroxyl group of Glu (Figure 72). However, as the amine group is not as strong a nucleophile as a hydroxyl group, the reaction between it and C1 happens slowly which causes a reduction of enzyme activity.

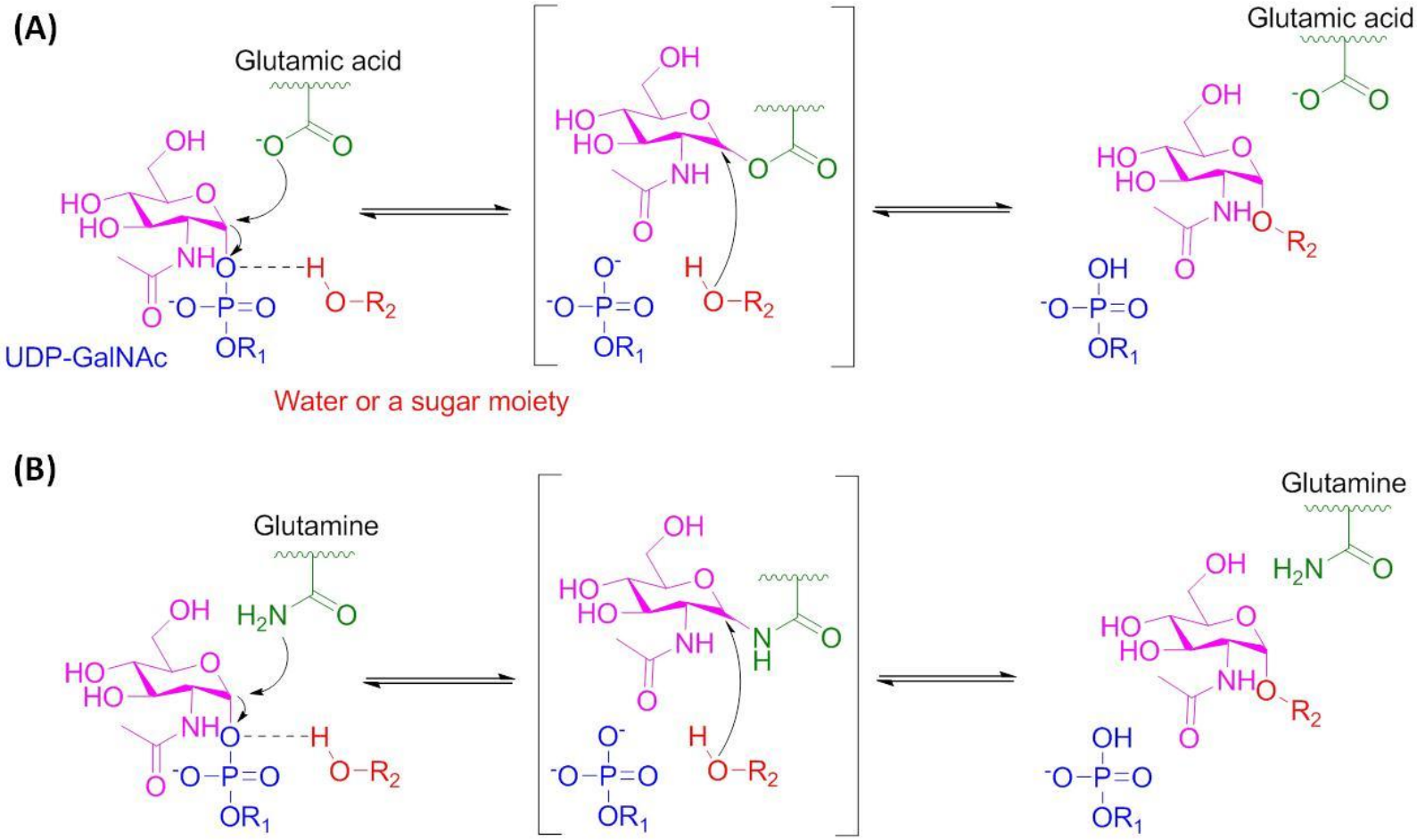


Figure 72. Diagram explains how the BoGT6a Glu192Gln retains enzyme activity. NH_2 group of Glu192 can partly play the role of OH group of Gln192 by giving/donating H atom to C1 of GalNAc moiety and form covalent bond. Diagram created using ChemDraw.

Nonetheless, it should be recognised that these are structures stabilised by being incorporated in different locations in the unusual monoclinic crystal lattice. Only structures A and B are present in molecules in the orthorhombic form. More structural and chemical studies are necessary to draw a conclusive evidence for BoGT6a catalytic mechanism.

3.3.3.4 Structure-function relationships in metal-dependent and metal-independent GT6

In the mammalian GT6, the metal-binding DXD motif is at the junction of the two subdomains and interacts directly with the ribose moiety of the UDP moiety as well as mediating metal ion interactions with the phosphate group of the UDP (Boix *et al.*, 2001, Patenaude *et al.*, 2002). The DXD motif is a shared feature of all GT families with GT-A folds with the exception of the metal-independent GT14 family (Breton *et al.*, 1998b, Breton and Imberty, 1999). However, in the bacterial enzymes, it is replaced by the NXN sequence except in GT6s from bacteriophage *Parachlamydia acanthamoebae*, and the cyanophage PSSM-2, which still retain the DXD motif and require metal ions for activity (Thiyagarajan *et al.*, 2012, Brew *et al.*, 2010). The new structures of BoGT6a E192Q in complex with the donor substrate show a small positional change in each residue of the NXN motif compared to that in each residue of the DXD motif (Figure 73).

In complexes containing free UDP, two cationic residues close to the C-terminus of α3GT, Lys359 and Arg365, interact with the diphosphate but in the complex of the low activity mutant of α3GT (Glu317Gln) with the substrate UDP-gal, the side chain of Lys359 is disordered and only Arg365 interacts with the α-phosphate (Tumbale *et al.*, 2008). The interaction with Arg365 is facilitated by structural changes in a loop containing Trp195, with which Arg365 forms a stacking interaction. In the structure of a complex of the inhibitory substrate analog, UDP-2F-gal, with the Arg365Lys mutant of α-3GT, the side chain of Lys359 points towards the β-phosphate of the inhibitor but the C-terminal 9 residues including Arg365 are disordered (Jamaluddin *et al.*, 2007). It should be noted that these complexes contain catalytically impaired mutants of the enzyme and, in one case, an inhibitor rather than substrate and are imperfect models of the enzyme-substrate complex. Both Lys359 and Arg365 make

contacts with UDP in its complex with α3GT (Jamaluddin *et al.*, 2007, Tumbale *et al.*, 2008) and conservative substitutions of Lys359 to Arg or Arg365 to Lys result in >30-fold reductions in k_{cat} but have small (~2-fold) effects on the K_M for UDP-gal (Jamaluddin *et al.*, 2007); the Lys to Ala mutation produces a 350-fold reduction in k_{cat}. Therefore Lys359 and Arg365 are important, although not essential, for activity and appear to have a principal role in transition state stabilisation, mediated through interactions with the UDP leaving group, as opposed to (ground-state) substrate binding.

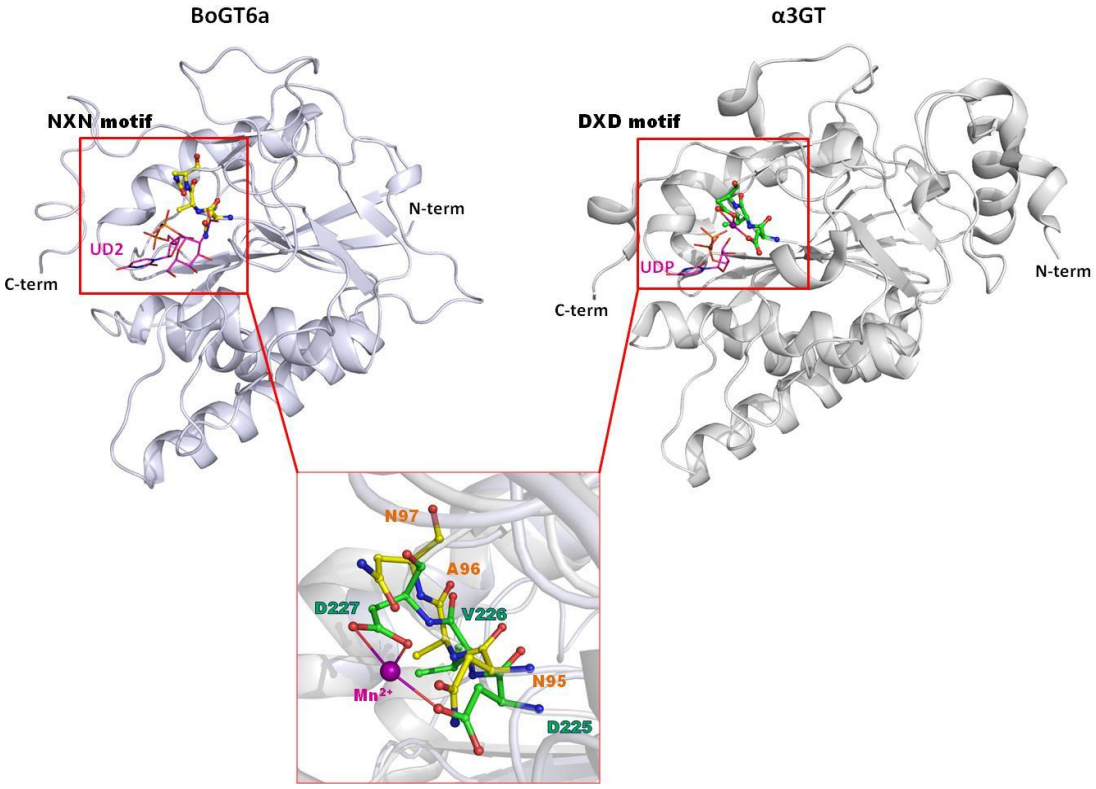


Figure 73. Structural comparison of the metal independent BoGT6a (with the NXN motif) and the metal dependent bovine α3GT (PDB 1K4V(Boix *et al.*, 2001)) (with the DXD motif). The protein are shown in the cartoon representation and coloured in silver for the BoGT6a and in grey for the bovine α3GT**.** The inset shows the details of the NXN motif (in yellow) and the DXD motif (in green). The residues were shown as stick, the UDP-GalNAc (noted as UD2) and UDP as lines and colour in magenta. Metal ion was shown as sphere in magenta. Picture created using Pymol.

Lys231 of BoGT6a is homologous with Lys359 of α3GT, a residue that is conserved in all metal-dependent (DXD) and metal-independent (NXN) GT6. However the

residue corresponding to Arg365 of α3GT is conserved in most other metal-dependent GT6 but not in the bacterial enzymes. The present structural studies indicate that Arg243 of BoGT6a has a similar structural location to α3GT Arg365 but appears to interact less specifically with the diphosphate moiety. Also, it is not within H-bonding distance of the diphosphate in structure C. Substitution of Lys231 by Ala increased the K_M about 2-fold but reduced k_{cat} more than 200-fold whereas the double mutation of Arg243 and Arg244 to Ala reduced k_{cat} by a factor of 10 (Tumbale and Brew, 2009). Thus, Lys231 has a greater stabilising effect on the transition state than ground state and appears to help to stabilise the UDP leaving group during catalysis, a role similar to that of the homologous Lys in α3GT that is consistent with the present structural studies.

Other metal-independent members of the GT-A superfamily- The majority of Leloir GTs, enzymes that utilise sugar nucleotides as donor substrates, group into either of two large superfamilies that have GT-A and GT-B folds. The GT-B superfamily is metal-independent whereas most representatives of the GT-A fold superfamily that have been functionally characterised are metal-dependent and have DXD metal-binding motifs (Lairson *et al.*, 2008). The sialyltransferases, whose donor substrate is a nucleotide monophosphate sugar (CMP-sialic acid) are metal-independent and include proteins with GT-A and GT-B-like folds, but with distinct topologies; comparisons of structure-function relationships between sialyltransferases and other GTs are challenging because of their non-standard folds and the character of the donor substrate (Audry *et al.*, 2011). Among the members of the "standard" GT-A group that have been functionally characterised, only the members of GT14 are metal-independent like BoGT6a. The GT14 and GT6 families differ in catalysing inverting and retaining reactions, respectively, but GT14 members also differ in having no conserved motif corresponding to DXD (or NXN). The crystallographic structures of one GT14, C2GnT-L, in the apo-form and in complexes with either acceptor substrate or UDP have been determined (Pak *et al.*, 2006, Pak *et al.*, 2011). The enzyme is a disulphide-bonded dimer and, in the complex with UDP, the two molecules of each dimer are in "open" and "closed" conformations. In the closed conformation, two basic residues close to the C-terminus of the protein, Arg378 and Lys401, interact with the β-phosphate of the UDP but these interactions are not

present in the open conformer. Substitution of either Arg378 or Lys401 by Ala eliminates catalytic activity but the Arg378 mutation has relatively small effects on the binding of UDP or UDP-GalNAc whereas the Lys401 mutation eliminates binding. Based on these observations, these two cationic residues were proposed to fulfil the role of the metal ion in metal-dependent GT-A GTs. Although characterised enzymes of the GT14 family are metal-independent, in some, the activity is enhanced by divalent metals (Sun *et al.*, 2007). Therefore they appear to have a metal-binding site and may have evolved from a metal-dependent ancestor. In contrast, the activity of BoGT6a decreases when increasing levels of Mn^{2+} ion are added to the enzyme (Tumbale and Brew, 2009).

3.4 Conclusions

In summary, the structures of BoGT6a in complex with either its acceptor substrate or the donor substrate show that the enzyme active site is built by residues close to both the N- and C-terminal regions and show that the C-terminally truncated BoGT6a, including all the key residues, represents the minimum size of a functional GT6.

Regardless of the modest resolutions, the presence of non-crystallographic symmetry provides us with structures that are snapshots of potential intermediates in the hydrolysis of UDP-GalNAc. The Gln192 mutation in BoGT6a greatly reduces but does not eliminate the glycosyltransferase and hydrolase activities. This is because the residue Gln with the amine group still can be a catalytic base residue although it is not as effective as the hydroxyl group of the residue Glu.

In addition, the presence of the configuration C in the BoGT6a E192Q•UDP-GalNAc form III structure suggests BoGT6a follows the double displacement mechanism to retain the α configuration of the GalNAc moiety from the donor substrate UDP-GalNAc.

As discussed above, two C-terminal basic residues in BoGT6a interact with the donor substrate, but these interactions are similar to those in metal-dependent GT6. If we surmise that the GT-A superfamily evolved from a metal-dependent common ancestor it would seem that the GT6 and GT14 families have used different

adaptations to become metal-independent. In the GT6 family we propose that the replacement of the DXD motif by NXN was a major factor in the transition between metal-dependence and metal-independence. This double substitution removes the requirement for a divalent metal ion to counter charge repulsion between the aspartates and diphosphate of the UDP. Therefore, in the metal-dependent GT6, the role of the metal ion in donor substrate binding in the ground state and stabilising the UDP leaving group in the transition state (Lairson *et al.*, 2008) is effectively performed by the polypeptide in the metal-independent GT6.

CHAPTER IV

Crystallisation of the BoGT6a E192Q in complex with FAL and UDP-GalNAc

4 Crystallisation of BoGT6a E192Q in complex with its acceptor (FAL) and donor (UDP-GalNAc) substrates

4.1 Methods

4.1.1 Expression of BoGT6a E192Q

4.1.1.1 Preparation of BoGT6a E192Q expression cell stock

BoGT6a E192Q cloned into the vector pET42(+) was kindly provided by our collaborator; Professor Keith Brew, Florida Atlantic University, USA. In order to generate sufficient copies of the plasmid for subsequent transformation into an expression strain, the recombinant plasmid, which was fixed on a filter paper, was dissolved in 10 µl of DNA/RNA free water and 5 µl of this solution was transformed into *E. coli* DH5α cells using the heat shock method. The recombinant *E. coli* DH5α was selected on Luria-Bertani (LB) agar media (Table 8) containing 50 µg/ml of Kanamycin. To confirm that the sequence was correct, plasmid, which was isolated from the recombinant *E. coli* DH5α using a Wizard Plus SV Minipreps DNA purification system (Promega), was sent to Eurofins for DNA sequencing with the universal T7 primer and T7 term primer.

Table 8. Ingredients of media used in BoGT6a E192Q expression

Media	Ingredients (per litre)
LB broth	10 g Tryptone 5 g Yeast extract 10 g NaCl
LB agar	10 g Tryptone 5 g Yeast extract 10 g NaCl 1.5 g Agar

E. coli BL21 CodonPlus (DE3)-RIPL competent cells were transformed with the plasmid isolated from *E. coli* DH5α using the heat shock transformation method. The transformed cells were spread onto LB agar media containing 50 µg/ml of Kanamycin and incubated at 37 ^{o}C overnight. One single colony was transferred to 10 ml of LB broth (Table 8) supplemented with 50 µg/ml of Kanamycin and the inoculation was incubated overnight at 37 ^{o}C with shaking at 250 rpm. Glycerol stocks were made by mixing 500 µl of the overnight cell culture with 500 µl of 20 %($^{v}/_{v}$) glycerol and stored at -80 ^{o}C for using in further BoGT6a E192Q protein expression.

4.1.1.2 Expression of BoGT6a E192Q

The frozen glycerol stock was used to inoculate 10 ml of LB broth containing 50 µg/ml of Kanamycin in a 50 ml falcon tube. The culture was allowed to grow at 37 ^{o}C with shaking at 250 rpm overnight. The whole overnight culture was used to inoculate 1 L of LB containing 50 µg/ml Kanamycin and incubated at 37 ^{o}C with shaking at 250 rpm until the OD_{600} reached 0.8 – 1.0. The temperature was then changed to 24 ^{o}C for expression overnight. After 20 hours the expression cultures were harvested by centrifugation at 4000 rpm for 20 min at 4 ^{o}C using a Beckman Coulter Avanti J-25 Centrifuge. The cell pellet was used for purification.

4.1.2 Purification of BoGT6a E192Q

4.1.2.1 Preparation of purification sample

The cell pellet was washed by resuspension in 50 ml of 25 %%($^{w}/_{v}$) sucrose, 20 mM Tris-HCl, pH 7.0, and then collected by centrifugation at 5000 rpm, 4 ^{o}C , for 40 min using a Thermo Scientific Heracus Megafuge 16R Centrifuge. The cell pellet was then resuspended in lysis buffer (Table 9) and the cells were lysed using the cell disrupter (Constant Systems) at 20 kpsi. The cell lysate was centrifuged at 25,000 rpm for 40 min at 4 ^{o}C using a Beckman Coulter Avanti J-25 Centrifuge. The supernatant was then used in subsequent affinity purification steps (called starting protein sample). To analyse how much of the protein was in insoluble form, 0.05 g of cell pellet was mixed with 200 µl lysis buffer, and the supernatant (called cell pellet sample) was collected after centrifugation at 13, 000 rpm, 4 ^{o}C for 1 min using

an Eppendorf Centrfuge 5415D. 16 µl of each sample were mixed with 4 µl of 5X SDS-PAGE loading dye. The mixtures were heated at 95 ^{o}C for 5 min and stored at 4 ^{o}C until they were analysed by Bis-tris SDS-PAGE.

4.1.2.2 Purification of BoGT6a E192Q

4.1.2.2.1 Affinity chromatography using Ni^{2+} column

The cell lysate supernatant was applied to a 5 ml His-trap column (GE), which had previously been equilibrated with 50 column volumes (CV) of equilibration buffer (Table 9) at a flow rate of 3 ml/min. The column was then washed with 10 CV of washing buffer 1 (Table 9) to eliminate nucleic acid contaminants followed by 20 CV of washing buffer 2 (Table 9) to remove impurities that were weakly to moderately bound to the column. Finally, the protein was eluted by application of the elution buffer (Table 9) and fractions were collected (1 ml/fraction).

All eluted fractions were placed on ice. 100 µl of each the cell lysate supernatant, loading flow through, washing step 1 flow through, washing step 2 flow through and the elution fractions were kept for analysis by Bis-Tris SDS-PAGE and Western blot. 16 µl of each sample were mixed with 4 µl of 5X SDS-PAGE loading dye. The mixtures were heated at 95 ^{o}C for 5 min and loaded onto a Bis-Tris SDS-PAGE gel. 15 % resolving, 4 % stacking Bis-Tris SDS-PAGE gel was run for all samples at 200 V at room temperature until the dye front migrated to the edge of the gel. The gel was stained with brilliant blue R-250 stain for an hour and washed with destaining solution until the protein bands were clearly visible.

For Western blot detection, all samples were prepared and loaded onto a Bis-Tris SDS-PAGE gel as the process of Bis-Tris SDS-PAGE. The gel was also run at 200 V at room temperature until the dye front migrated to the edge of the gel. After that, the samples were transferred to a PVDF membrane (Millipore) using a transfer cassette (Biorad). The transfer was set up at 60 V in 1 hour. The membrane was then incubated in a blocking solution (5% milk powder in TBST (20 mM Tris-Cl pH 7.5, 0.15 M NaCl and 0.1% Tween 20)) in 1 hour to prevent the interactions between the membrane and the antibody used for detection of the target protein. The membrane was washed three times for 10 min each with TBST solution before being incubated

with a monoclonal anti-polyHistidine conjugated with a horseradish peroxidase (HRP) antibody (code A7058-1VL, Sigma) for 1 hour. The proteins were detected by incubating the membrane in 10 ml of a colourimetric detection solution (0.1% 3,3'-diaminobenzidine tetrahydrochloride (DAB), 0.002% hydrogen peroxide in TBST) until the blots appeared.

4.1.2.2.2 Size exclusion chromatography

Size exclusion chromatography was performed to enhance the purity of the protein. Protein eluted from the His-trap column was concentrated by centrifugation at 4000 rpm, 4 ^{o}C using Amicon Ultra-15 MW3000 spin concentrators (Millipore) to get 1 ml of protein solution in the storage buffer. Concentrated protein was loaded onto a Superdex 200 16/60 column (GE) that had been equilibrated with 1.5 CV of storage buffer (Table 9). Peak fractions were collected and analysed by gel electrophoresis and Western blotting as described previously.

Table 9. Ingredients of buffers used in purification of BoGT6a E192Q

Buffer	Ingredients
Lysis buffer	20 mM Tris-HCl, 1 mM EDTA, pH 8.0
Equilibration buffer	20 mM Tris-HCl, pH 8.0, 100 mM NaCl
Washing buffer 1	20 mM Tris-HCl, pH 7.0, 100 mM NaCl, 5 mM Imidazole pH 8.0
Washing buffer 2	20 mM Tris-HCl, pH 8.0, 100 mM NaCl, 60 mM Imidazole pH 8.0
Elution buffer 1	20 mM Tris-HCl, pH 8.0, 100 mM NaCl, and 500 mM Imidazole pH 8.0
Storage buffer	20 mM Tris-HCl, pH 8.0, 100 mM NaCl, 2 mM DTT

Fractions containing BoGT6a E192Q were concentrated to a final concentration of 8 mg/ml by centrifugation at 4000 rpm, 4 ^{o}C with Amicon Ultra-15 MW3000 spin concentrators (Millipore) for use in crystallisation. A sample of the purified protein was also transferred into water and the molecular weight analysed by electrospray mass spectrometry.

4.1.3 Crystallisation of BoGT6a E192Q in complex with its ligands

BoGT6a E192Q obtained from the purification was used directly for crystallisation. The complex of the enzyme and its donor, UDP-GalNAc, was set-up by adding 100 mM UDP-GalNAc to the protein solution such that the final concentration was 10 mM. The complex of the enzyme with both its donor and acceptor was formed by adding 100 mM FAL to the mixture of protein and UDP-GalNAc (10 mM) to a final concentration of 10 mM. In an attempt to trap the intermediate state, these complexes were crystallised immediately after setting up. All processes were performed at room temperature.

Potential crystallisation conditions for the BoGT6a E192Q·UDP-GalNAc complex were screened using the Phoenix crystallisation robot. Two commercial crystallisation screens: Structure Screen 1 & 2 and Proplex (Molecular Dimensions Ltd), were set-up in 96-well Intelli-plate® (Art Robbins Instrument). 0.2 µl sitting drops were set with a 1:1 protein: reservoir ratio. All of the plates were incubated at 16 ^{o}C. Based on the hits obtained from this crystallisation screening, optimisations were performed in 24-well plates by varying the precipitant concentration but maintaining a constant ratio of protein: reservoir solution and incubation temperature.

Co-crystallisations of BoGT6a E192Q with both UDP-GalNAc and FAL were set up based on crystallisation hits for the BoGT6a E192Q·UDP-GalNAc complex in 24-well plates. The plates were incubated at 16 ^{o}C. BoGT6a E192Q·UDP-GalNAc complex crystals were also soaked with 100 mM FAL in an attempt to obtain crystals of BoGT6a in complex with both its donor and acceptor substrates.

4.2 Results

4.2.1 Expression and purification

The plasmid pET42(+)-BoGT6a E192Q received from our collaborator was propagated in E. *coli* DH5α cells and verified by DNA sequencing. Since this is a long sequence (738 bps without His-tag), the sequencing was performed with both T7 primer and T7 term primer. The combined results displayed the sequence of the full-length of the enzyme with 738 bps and 57 bps of the His-tag. The sequencing results were translated to give the 265 residue protein sequence. The mutation E192Q (GAA to CAA) was confirmed along with the presence of 19 residues that form the His-tag at the N-terminus of the protein (Figure 74). The protein mass calculated by using ExPASy – Compute pI/Mw (Expert Protein Analysis System) tool (Gasteiger *et al.*, 2003) is 31033 Da and its pI is 7.86.

The expression protocol used for BoGT6a E192Q was adapted from the native BoGT6a expression protocol established by Tumbale *et. al* (Tumbale and Brew, 2009). Although the transformed *E. coli* BL21 CodonPlus (DE3)-RIPL cells were expected to express BoGT6a E192Q in soluble form by leaky expression, gel electrophoresis analysis showed that some protein was present in the cell pellet (Figure 76). However this was not a significant amount and so the insoluble protein was not used for purification, as this would have required modification of the purification protocol.

DNA sequence

atg ggc agc agc cat cat cat cat cat cac agc agc ggc ctg gtg ccg cgc ggc agc cat **atg** aga att ggt ata tta tat atc tgt act ggc aaa tat gac att ttt tgg aaa gac ttt tat cta agc gca gaa cgt tat ttt atg caa gac caa tct ttc att atc gag tat tat gta ttt act gat agt cct aaa cta tat gac gaa gaa aac aac aaa cat att cac cgg atc aaa caa aag aat tta gga tgg cct gac aac aca tta aaa cgt ttc cat ata ttc ctt cgt atc aag gaa cag tta gag cga gaa acc gac tat cta ttt ttc ttc aat gcc aat ctc tta ttc acc agt cct att ggc aaa gaa att cta cca cca tca gat agt aac gga tta cta gga act atg cac cct gga ttc tac aat aaa ccg aac tcc gaa ttt aca tac gag cga aga gat gct tct act gcc tat atc cca gag gga gaa ggt cga tat tat tac gct gga ggg ctt tca ggt gga tgt aca aag gcc tac ttg aaa ctc tgc aca aca att tgc tca tgg gtt gac aga gat gcc aca aac cat ata ata cca att tgg cac gac **caa** tct cta atc aat aaa tac ttt tta gat aat cca cca gct att aca ttg tcc cct gca tat cta tac cca gaa ggt tgg ctc ctt cct ttt gaa cca ata atc ctc att cga gac aaa aat aaa ccc caa tat ggc ggg cat gaa tta ttg cga aga aaa aac **tga**

Protein sequence

M G S S H H H H H H S S G L V P R G S H **M** R I G I L Y I C T G K Y D I F W K D F Y L S A E R Y F M Q D Q S F I I E Y Y V F T D S P K L Y D E E N N K H I H R I K Q K N L G W P D N T L K R F H I F L R I K E Q L E R E T D Y L F F F N A N L L F T S P I G K E I L P P S D S N G L L G T M H P G F Y N K P N S E F T Y E R R D A S T A Y I P E G E G R Y Y Y A G G L S G G C T K A Y L K L C T T I C S W V D R D A T N H I I P I W H D **Q** S L I N K Y F L D N P P A I T L S P A Y L Y P E G W L L P F E P I I L I R D K N K P Q Y G G H E L L R R K N **Stop**

Figure 74. DNA and amino acid sequences of BoGT6a E192Q. The sequence of the His-tag was highlighted in orange and the E192Q mutant sequence was bold in red.

As a His-tag fused protein, BoGT6a E192Q was purified by the nickel affinity chromatography method using a 5 ml His-trap HP column (GE). The first trial purification was adapted from the purification protocol of native BoGT6a, in which contaminants were washed out through two wash-steps at 5 and 60 mM imidazole. The target protein was eluted at 500 mM imidazole (Tumbale and Brew, 2009). The chromatogram showed a high absorbance peak during the sample loading step (2200 mAu), a small peak at the second washing step (600 mAu) and a high peak at the elution step (1000 mAu) (Figure 75). Although the flow through collected during sample loading and the washing fractions contain a high concentration of protein, western blot results indicated that there was no BoGT6A E192Q present in the washing fractions and only a small amount present in the flow through fraction (Figure 76).

6 ml of the elution fraction at a concentration of 2 mg/ml was dialysed in order to transfer it into storage buffer and concentrated to 8 mg/ml. The final yield of BoGT6a E192Q was 12 mg of soluble protein per 1 L of cell culture. This is similar to the yield of native BoGT6a reported by Tumbale *et. al* (2009). Gel electrophoresis result showed 2 bands around 27 kDa in the elution fraction, but only one of these was present on the western blot (Figure 76). The lower molecular weight band on the protein gel was comparable to the theoretical molecular weight of the protein and was also reported in the purification of native BoGT6a (Tumbale and Brew, 2009). There was a smear associated with the higher band that may be contaminants from the cell culture. Thus further purification steps were required.

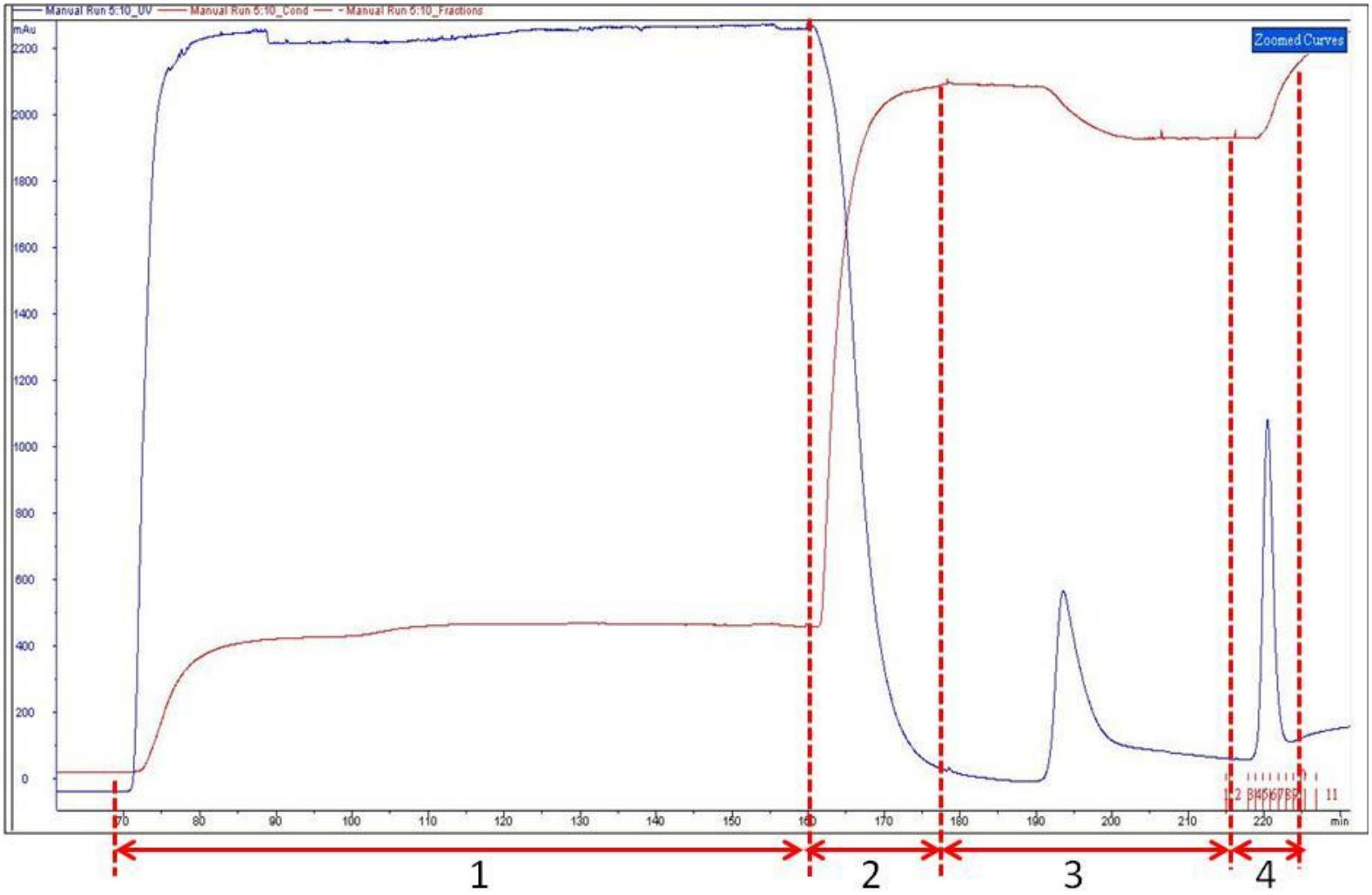


Figure 75. Chromatography of the first trial affinity purification. (1) flowthrough fraction, (2) wash 1 step fraction, (3) wash 2 step fraction, and (4) elution fraction. The blue line indicates the UV intensity (mAu) and the pink line the conductivity of the solutions.

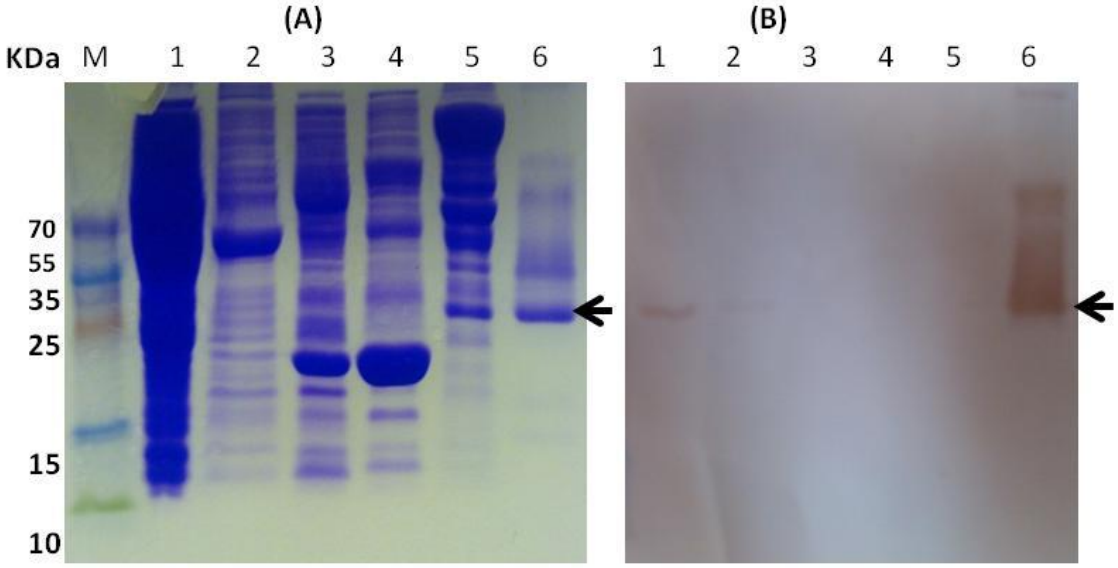


Figure 76. Bis Tris SDS-PAGE analysis of first trial purification of BoGT6a E192Q. (A) shows the Bis Tris SDS-PAGE result and (B) the Western blot result. (M) 10 µl of Thermo Scientific PageRuler Plus Prestained protein ladder, (1) supernatant of cell lysate, (2) supernatant of the cell pellet after lysis, (3) flow through fraction, (4) wash 1 fraction, (5) wash 2 fraction, and (6) elution fraction. 20 µl of each sample was loaded.

In an attempt to obtain a higher purity of BoGT6a E192Q, a lower elution gradient was applied with elution buffer 2, but the purity of the elution fraction was not improved. Size exclusion chromatography using Superdex 200 resin (GE) was thus used to separate the target protein from the impurities. There were three peaks appearing on the size exclusion chromatogram (Figure 77). A comparison of the retention time of those peaks and a standard graph shows that the first peak with a retention time about 85 ml corresponds to 44 kDa of protein weight.The second peak corresponds to 33 kDa of protein weight, which is the expected weight of the target protein. The last peak is too small comparing to the standard graph which can be the absorbance of imidazole in the buffer or contaminant from the cell culture. All of the fractions in each peak were pooled together and analysed using gel electrophoresis and Western blot.

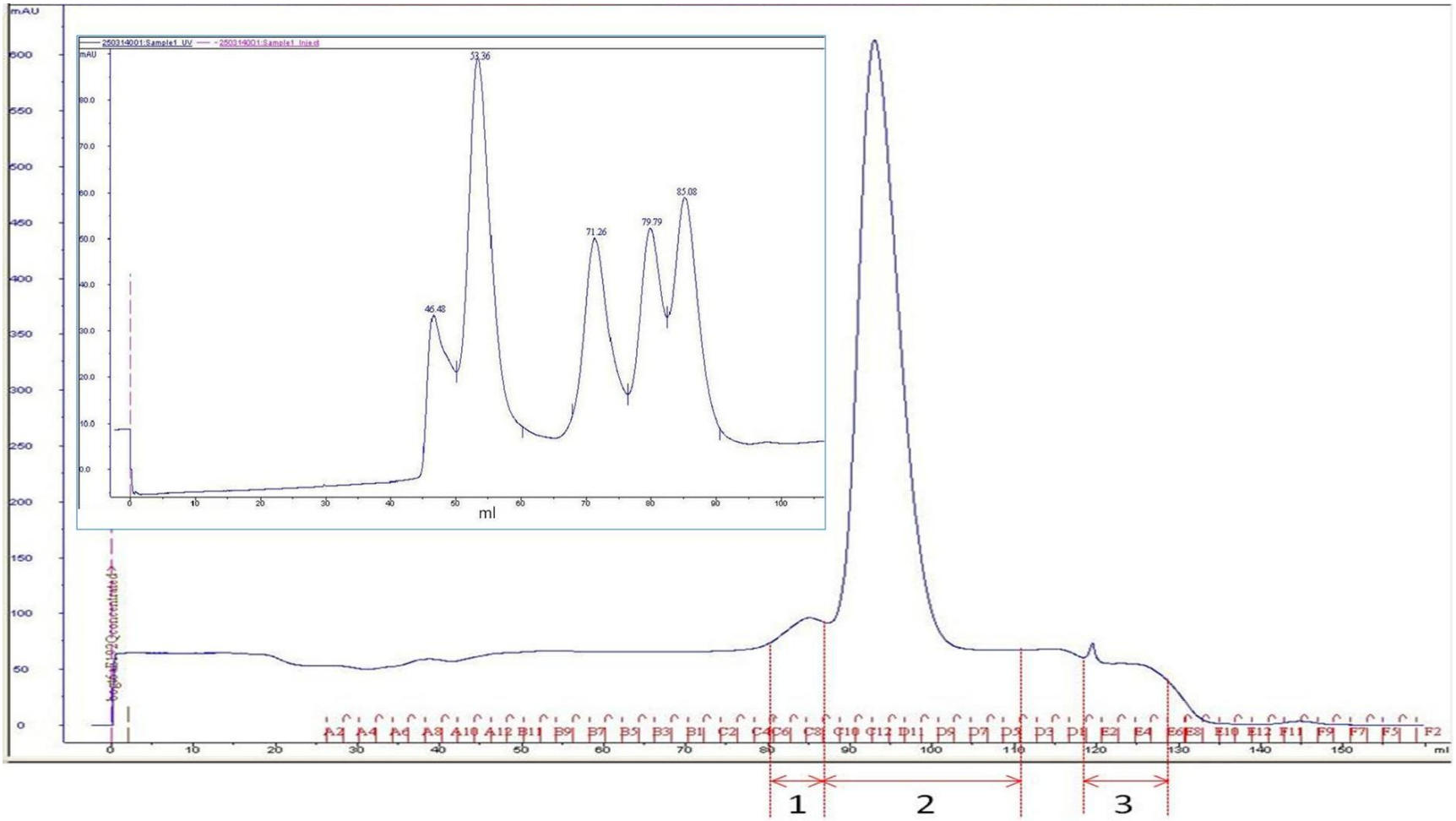


Figure 77. Chromatograph of the BoGT6a E192Q purification using size exclusion chromatography method. The collected fractions are marked in which 1 included C5-C8, 2 included C9-D5, and 3 included E1-E5. The inset is the standard graph in which the retention time of each peak is marked. The blue line indicates the UV intensity (mAu).

On the gel electrophoresis result, peak 1 showed only one band near 27 KDa, the expected molecular weight of BoGT6a E192Q on the Bis Tris SDS-PAGE analysis. The peak 2 had two bands, with the major band also at the expected position for the protein. Peak 3 showed no bands (Figure 78). This was either because there was insufficient protein to be detected on the gel or the absorbance was from the imidazole in the sample and not from protein. This was confirmed by Western blotting (Figure 78). The fractions containing BoGT6a E192Q were pooled together and concentrated by centrifugation with Amicon Ultra-15 MW3000 (Millipore) to a final concentration of 8 mg/ml for use in crystallisation.

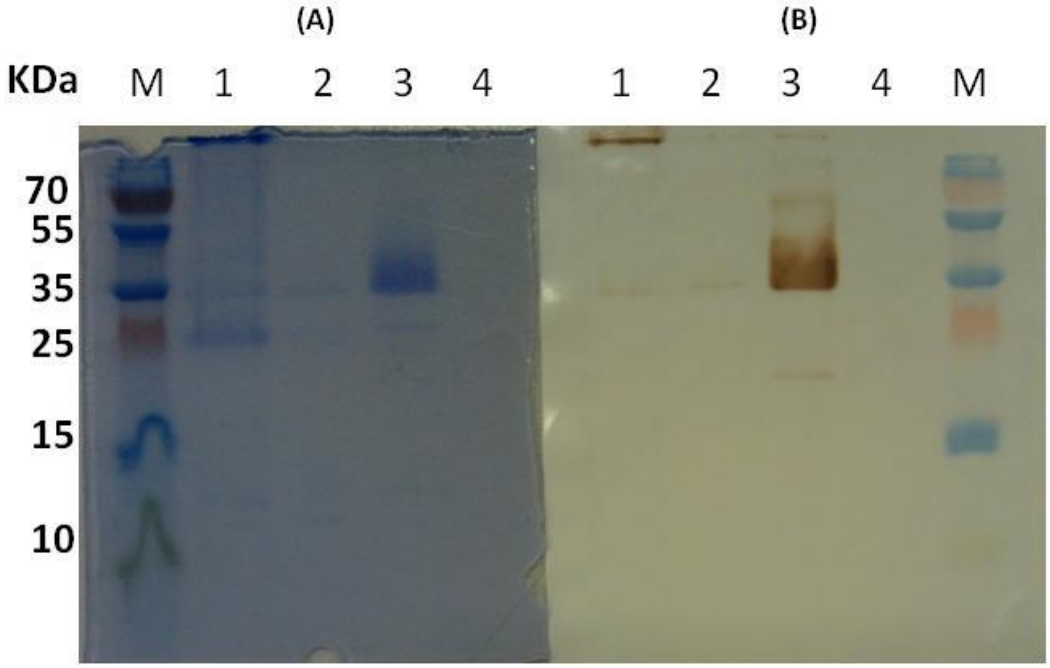


Figure 78. Gel electrophoresis and Western blot results of analysing the BoGT6a E192Q purity after the size exclusion purification. (A) Gel electrophoresis result and (B) Western blot result. (1) 5 µl of the protein before applying to the Superdex 200 16/60 column (GE), (2) C5-C8 fraction, (3) C9-D5 fraction, (4) E1-E5 fraction, and (M) 10 µl of Thermo Scientific PageRuler Plus Prestained protein ladder. 20 µl of each fraction was loaded on gel.

The purified protein was analysed by MS and the result showed one peak with a mass of 31033.45 Da (Figure 79). The single strong peak from MS indicated that the protein solution was pure enough for crystallisation. Although the band of BoGT6a E192Q on the gel electrophoresis result appeared to be positioned around 27 kDa, the mass from MS corresponded to theoretical molecular weight of the protein,

31034 Da, calculated by using ExPASy – Compute pI/Mw (Expert Protein Analysis System) tool (Gasteiger *et al.*, 2003). The molecular weight of the protein was also in agreement with the DNA sequencing results that showed the presence of the long His-tag at the N-terminal region, and also corresponded to the form III structure of BoGT6a E192Q in complex with UDP-GalNAc in which first three His-tag residues were visible. With a successful expression and purification protocol established for BoGT6a E192Q the protein could be supplied consistently in the same condition and at sufficient purity for crystallisation. This facilitated the searching for crystallisation conditions for the BoGT6a E192Q complexes.

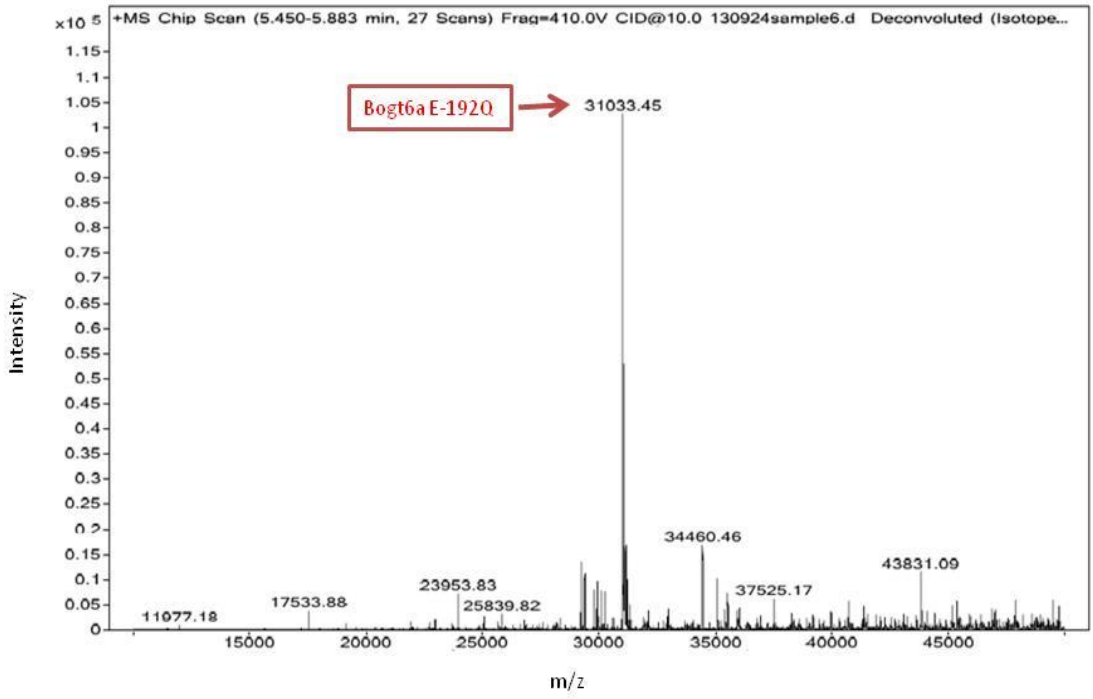


Figure 79. The mass spectrometry result for BoGT6a E192Q.

4.2.2 Crystallisation of BoGT6a E192Q in complex with its ligands

BoGT6a E192Q in storage buffer was concentrated to 8 mg/ml by centrifugation at 4000 rpm, 4 ^{o}C using an Amicon Ultra-15 MW3000 (Millipore). The concentrated protein solution was used to set-up two commercial crystallisation screens: Proplex and Structure Screen 1 & 2. However, only the Proplex screening gave hits (Figure 80). Crystals appeared in wells D4, E11 and E12 after only two days, whilst in other wells it took two weeks for crystals to form. Conditions D4 and E12 were repeated on 24 well plates with the same protein concentration and incubation condition. Only

small crystals appeared in the D4 condition while bigger crystals were obtained from the E12 condition, however, crystals from both conditions were too fragile to be mounted.

The conditions were optimised by replacing $(NH_4)_2SO_4$ with Li_2SO_4, and screening different PEG 8000 concentrations (20 % – 18 % – 15 %). The condition comprising 0.1 M MES pH 6.5, 15 % (w/v) PEG 8000, and 0.2 M Li_2SO_4 gave bigger and better crystals. Crystals of BoGT6a E192Q•UDP-GalNAc from this condition were soaked with 10 mM FAL and left for 1 day at 16 ^oC in an attempt to obtain the ternary complex. This condition was also used for co-crystallisation of BoGT6a E192Q in complex with the acceptor, FAL, and the donor, UDP-GalNAc. This also yielded many bar-shaped crystals.

Table 10. Conditions of "hits" for crystallisation of BoGT6a E192Q in complex with UDP-GalNAc using the ProPlex Screen HT-96

Well	Condition		
	Salt	Buffer	Precipitant (w/v)
B12	0.1 M magnesium chloride	0.1 M Na HEPES pH 7.0	15 % PEG 4000
C11		0.1 M sodium cacodylate pH 6.5	25 % PEG 4000
D3	0.1 M potassium chloride	0.1 M Na HEPES pH 7.0	15 % PEG 5000 MME
D4	0.2 M ammonium sulphate	0.1 M Tris pH 7.5	20 % PEG 5000 MME
E11		0.1 M sodium citrate pH 5.0	20 % PEG 8000
E12	0.2 M ammonium sulphate	0.1 M MES pH 6.5	20 % PEG 8000

Some of the crystals from these crystallisation experiments were analysed at Diamond Light Source station I04. Since PEG 8000 was used, the crystals were only

soaked with the reservoir solution as a cryo-protectant before being cooling in liquid Nitrogen. The data collection process was not successful because there was ice covering some of the crystals and the maximum resolution of the diffraction data was only 4.5 Å (Figure 82). Although there were two datasets collected, these datasets are insufficient to determine the complex structure due to its low resolution (Table 11).

Nevertheless, these results indicated that the crystals obtained were indeed protein crystals. This is a promising result because the cryo-protection process can be optimised. Due to time constraints, no further crystallisation optimisations were performed, but as this is the only crystallisation condition so far to have yielded reproducible crystals, its discovery is a potential step towards BoGT6a crystallisation.

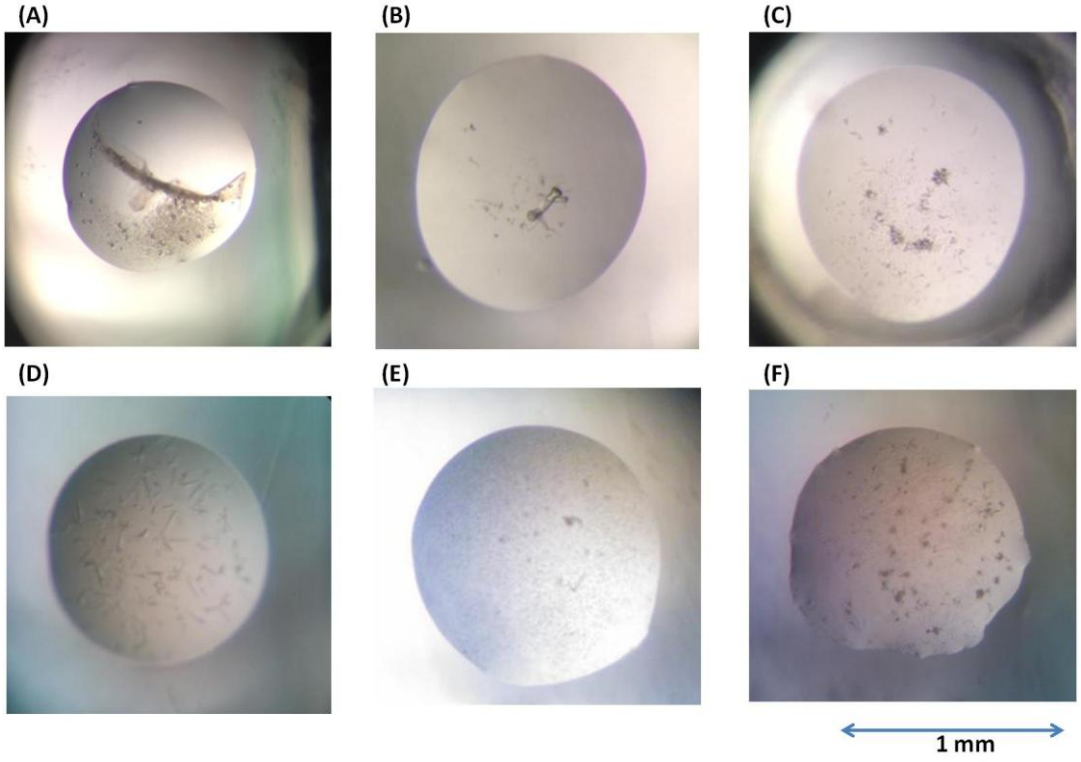


Figure 80. Crystals of BoGT6a E192Q in complex with UDP-GalNAc obtained from "hit" conditions. (A) from well B12, (B) from well C11, (C) from well D3, (D) from well D4, (E) from well E11 and (F) from well E12.

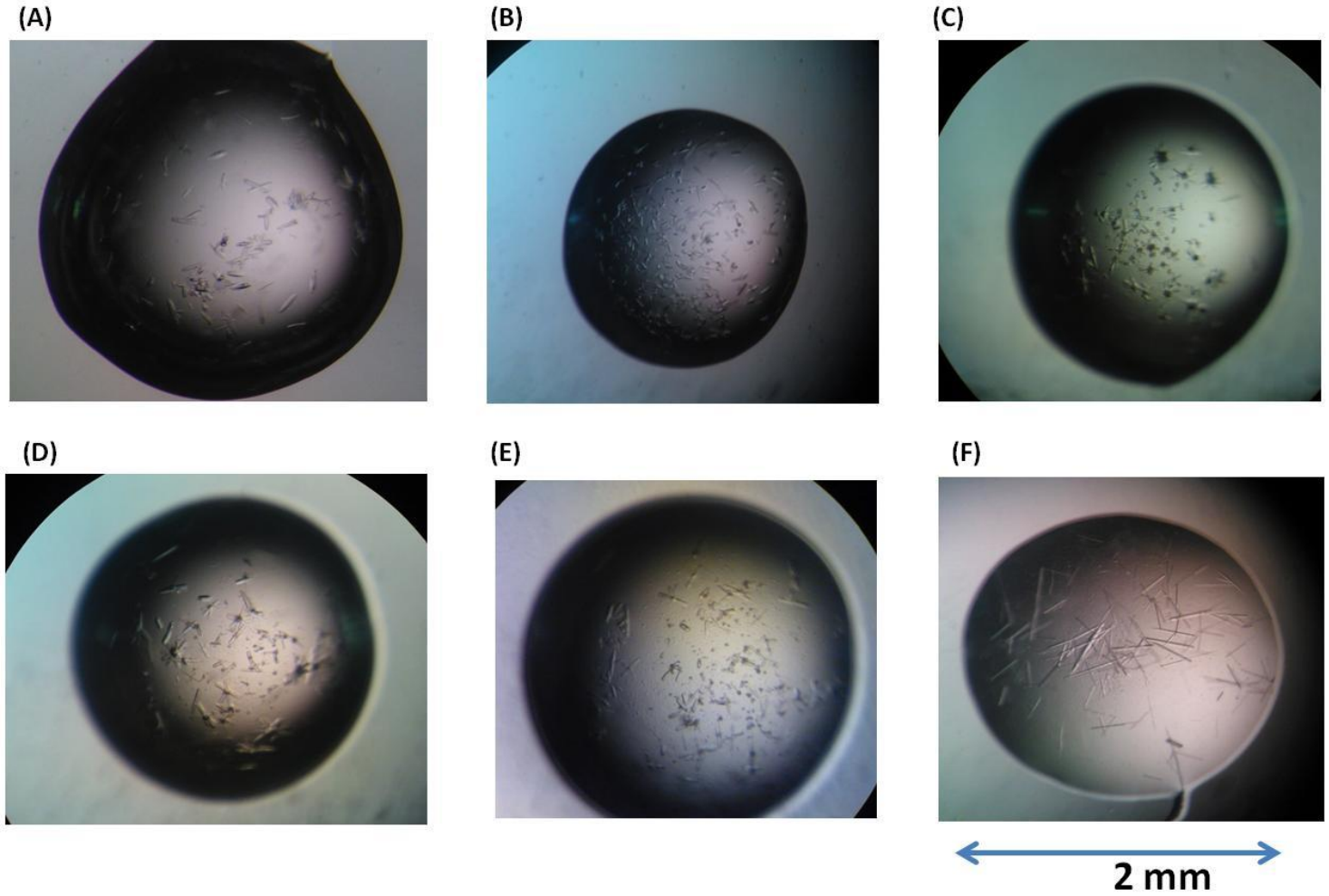


Figure 81. Crystals of BoGT6a E192Q in complex with UDP-GalNAc (A, B C and D) or with both UDP-GalNAc and FAL (E, and F). (A) the D4 condition, (B) the E12 condition, (C) the E12 condition in which Li_2SO_4 was used instead of $(NH_4)_2SO_4$, (D) and (E) the same condition as (C) before and after FAL was added respectively. (F) BoGT6a E192Q•UDP-GalNAc•FAL crystal in the condition as (C).

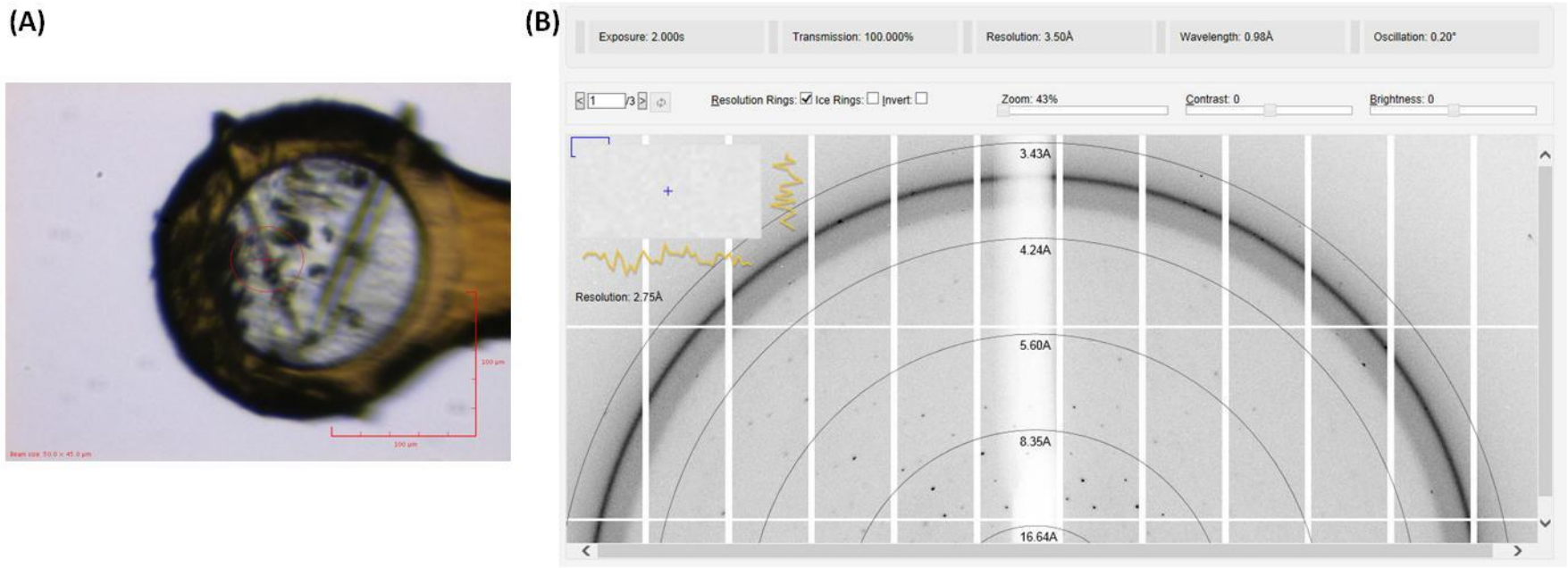


Figure 82. Diffraction data from a BoGT6a E192Q in complex with UDP-GalNAc and FAL crystal obtained from the condition of 0.1 M MES pH 6.5, 15% (w/v) PEG 8000, and 0.2 M Li_2SO_4.

Table 11. Information of data collections for BoGT6a E192Q in complex with UDP-GalNAc and FAL crystals

Dataset	Number of images	Resolution (Å)	Space group	Cell dimensions	R_{merge}	Completeness (%)
9	150	4.66	$P2_12_12$	a = 120.5 Å, b = 132.5 Å, c = 80.3 Å α = β = γ = 90.0 ^{o}	0.095	99.6
10	500	4.14	$P222_1$	a = 131.6 Å, b = 80.3 Å, c = 119.8 Å α = β = γ = 90.0 ^{o}	0.145	68.2

CONCLUSIONS AND FUTURE WORK

5 Conclusions and Future work

The glycosyltransferase family 6 (GT6) consists of Histo-blood group A and B glycosyltransferases (GTA, GTB), α-galactosyltransferase (α3GT), Forssman glycolipid synthase (FS) and Isogloboside 3 synthase (iGb3S). This family is of medical importance because their products affect the human immune system; the most well-known examples of which are GTA and GTB, which determine our blood group types. The overall structures of these enzymes are well conserved but small changes in the structures are important for the observed differences in their catalytic activity. In that context, exploring their structures and an understanding on how they link to the catalytic mechanism are important. This will also help in engineering these enzymes to synthesise non-natural glycoconjugates. *Bacteroides ovatus* glycosyltransferase 6a (BoGT6a) is a unique member of the GT6 family because of its NXN motif in its sequence, which replaces the well-conserved DXD motif observed in all vertebrate GT6. In mammalian GT6, the DXD motif is involved in a metal-dependent catalytic mechanism. The substitution raises a question about the effect of NXN motif in BoGT6a catalytic activity. In fact, the enzyme does not require metal ion for its catalytic activity (Tumbale and Brew, 2009). However, the structure of BoGT6a apo form shows a remarkably high structural similarity to its mammalian homologues such as GTA, GTB and α3GT, apart from its shorter N-terminal region (Thiyagarajan *et al.*, 2012).

To explore more about the enzyme catalytic mechanism, structural studies of the enzyme in complex with its donor substrate UDP-GalNAc and acceptor FAL were performed.

The crystal structure of BoGT6a in complex with FAL was obtained at 3 Å in the space group $P2_1$. The structure contains 4 molecules in the asymmetric unit, each of which has one FAL moiety bound in the acceptor binding site. A comparison of the BoGT6a apo form structure and the BoGT6a•FAL shows the significant conformational changes of the enzyme associated with acceptor binding, including the internal loop (residues Tyr126 to Arg151) which is absent in the enzyme apo form structure due to its high flexibility, the LBR-F (residues Trp189 to Glu192) which is involved in the enzyme activity and the C terminus. The conformational change of the C terminus which is known as a closed conformation upon ligand

binding and catalysis is also reported for most glycosyltransferase, including the other GT6 members (Lairson *et al.*, 2008, Boix *et al.*, 2001, Qasba *et al.*, 2005, Patenaude *et al.*, 2002). The acceptor binding site is highly conserved among GT6 family consisting of Glu192, Trp189, Thr134, Tyr153 and His122 (BoGT6a numbering) regardless their metal dependence or metal independence property. This suggests that metal ion only involves in the donor binding activity. This is consistent with the proposed role of the metal ion role in glycosyltransferase catalytic activity which electrostatically stabilise the developing negative charge of the nucleoside diphosphate leaving group (Lairson *et al.*, 2008).

Belonging to the retaining glycosyltransferase group, the catalytic mechanism of BoGT6a has not been clearly understood. There are two proposed mechanisms for BoGT6a, the double displacement and internal return (S_Ni-like) mechanisms. The difference between the two mechanisms is the presence of an intermediate stage in which a covalent bond is formed between the catalytic residue of the enzyme and the sugar moiety from the donor substrate. The potential catalytic base residue of the BoGT6a is Glu192. This residue is highly conserved in GT6 family. Kinetic assay of the mutant Glu192Gln (E192Q) shows a significant reduction (about 30000 fold) in the glycosyltransferase activity of BoGT6a. Structural study of this mutant in complex with its donor UDP-GalNAc is thus necessary to elucidate the enzyme catalytic mechanism.

There are three structures obtained for the BoGT6a E19Q•UDP-GalNAc complex. The form I structure was solved at 2.78 Å in the space group $P2_12_12_1$. This structure contains 4 molecules in the asymmetric unit. Each molecule has only one α-GalNAc in their active site. The second structure, called the form II structure, was also solved in the space group $P2_12_12_1$, but at a lower resolution (3.42 Å). This structure has 4 molecules in the asymmetric unit and has two configurations of the bound ligands. The configuration A is an intact UDP-GalNAc and the configuration B is a separate UDP and α-GalNAc. The final structure, which consists of 16 molecules in the asymmetric, was solved at 3.50 Å in the space group $P2_1$. This structure not only has the two configurations of the ligands observed in the form II structure but also has the configuration C which consists of UDP and β-GalNAc. The interesting feature of this configuration is that the β-GalNAc is in a close contact with the residue Gln192.

Such a close distance suggests a covalent bond formation between β-GalNAc and the Gln192. The configurations A, B and C are models for complexes in the hydrolysis reaction catalysed by BoGT6a, in which A is the substrate binding stage, C the short-lived intermediate stage and B the product release stage.

All three structures have provided the structural snapshots of BoGT6a catalysing UDP-GalNAc. The interactions between the enzyme and the donor substrate also gave an insight into the mechanistic role of the NXN motif cooperating with the residue Lys231 in the metal-independent activity of the BoGT6a. These structures together demonstrate how a significant divergence in catalytic properties can be accommodated by minor structural adjustments and explain the role of the NXN motif in BoGT6a metal independent catalytic activity.

In the form III structure, a link between residue Gln192 and C1 of β-GalNAc was built, based on the electron density between them. This suggests that BoGT6a may follow a double displacement mechanism. However this remains ambiguous in the current modest resolution structures, hence a higher resolution structure of BoGT6a E192Q in complex with UDP-GalNAc is now required to provide more information. From such a high resolution structure we might hope to determine the configuration, α or β, and orientation of the sugar product, and positions of water molecules in the active site.

Besides structural studies, more chemical evidence is required if we are to draw a conclusion regarding the enzyme mechanism. In the first instance, the existence of the link between residue Gln192 of the enzyme and C1 of β-GalNAc must be confirmed by other complementary methods, such as MS. MS experiments could be used to analyse the mixture of BoGT6a E192Q with the donor substrate, UDP-GalNAc, and/or the acceptor substrate, FAL. The slow reaction rate of BoGT6a E192Q may help to trap the intermediate stage, which would be indicated by the presence of a peak with the weight equal to that of the glycosyl-enzyme complex. Another way to prove the enzyme mechanism might be by incorporating deuterium into the product. If the enzyme follows the double displacement mechanism, a deuterium atom from environmental D_2O would replace the hydrogen of the amine group of Gln192. If this did not happen it would mean that the enzyme follows the

S_Ni mechanism; the GalNAc moiety of UDP-GalNAc interacts directly with the D_2O from the environment and the H atom of the Gln192 remains intact.

In addition, obtaining the ternary structure may help us to understand more about the role of BoGT6a in catalysing the transfer of GalNAc moiety from UDP-GalNAc to FAL. In the future, crystallisation conditions for BoGT6a E192Q in complex with both the donor substrate, UDP-GalNAc, and the acceptor substrate, FAL, should be optimised to produce better quality crystals. The recent crystallisation result seems to provide a good basis for production of superior crystals, as well as for the crystallisation of other mutants.

REFERENCES

References

ADAMS, P. D., AFONINE, P. V., BUNKOCZI, G., CHEN, V. B., DAVIS, I. W., ECHOLS, N., HEADD, J. J., HUNG, L.-W., KAPRAL, G. J., GROSSE-KUNSTLEVE, R. W., MCCOY, A. J., MORIARTY, N. W., OEFFNER, R., READ, R. J., RICHARDSON, D. C., RICHARDSON, J. S., TERWILLIGER, T. C. & ZWART, P. H. 2010. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D,* 66**,** 213-221.

ALFARO, J. A., ZHENG, R. B., PERSSON, M., LETTS, J. A., POLAKOWSKI, R., BAI, Y., BORISOVA, S. N., SETO, N. O., LOWARY, T. L., PALCIC, M. M. & EVANS, S. V. 2008. ABO(H) blood group A and B glycosyltransferases recognize substrate via specific conformational changes. *J Biol Chem,* 283**,** 10097-108.

AUDRY, M., JEANNEAU, C., IMBERTY, A., HARDUIN-LEPERS, A., DELANNOY, P. & BRETON, C. 2011. Current trends in the structure-activity relationships of sialyltransferases. *Glycobiology,* 21**,** 716-26.

BERGFORS, T. 2003. Seeds to crystals. *Journal of Structural Biology,* 142**,** 66-76.

BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N. & BOURNE, P. E. 2000. The Protein Data Bank. *Nucleic Acids Research,* 28**,** 235-242.

BOIX, E., SWAMINATHAN, G. J., ZHANG, Y., NATESH, R., BREW, K. & ACHARYA, K. R. 2001. Structure of UDP complex of UDP-galactose:β-galactoside-α-1,3-galactosyltransferase at 1.53-Å resolution reveals a conformational change in the catalytically important C terminus. *J Mol Biol,* 276**,** 48608-48614.

BOIX, E., ZHANG, Y., SWAMINATHAN, G. J., BREW, K. & ACHARYA, K. R. 2002. Structural basis of ordered binding of donor and acceptor substrates to the retaining glycosyltransferase, α-1,3-galactosyltransferase. *J Mol Biol,* 277**,** 28310-28318.

BRETON, C., BETTLER, E., JOZIASSE, D. H., GEREMIA, R. A. & IMBERTY, A. 1998a. Sequence-function relationships of prokaryotic and eukaryotic galactosyltransferases. *Journal of Biochemistry,* 123**,** 1000-1009.

BRETON, C., FOURNEL-GIGLEUX, S. & PALCIC, M. M. 2012. Recent structures, evolution and mechanisms of glycosyltransferases. *Curr Opin Struct Biol,* 22**,** 540-549.

BRETON, C. & IMBERTY, A. 1999. Structure/function studies of glycosyltransferases. *Curr Opin Struct Biol,* 9**,** 563-571.

BRETON, C., ORIOL, R. & IMBERTY, A. 1998b. Conserved structural features in eukaryotic and prokaryotic fucosyltransferases. *Glycobiology,* 8**,** 87-94.

BRETON, C., ŠNAJDROVÁ, L., JEANNEAU, C., KOČA, J. & IMBERTY, A. 2006. Structures and mechanisms of glycosyltransferases. *Glycobiology,* 16**,** 29R-37R.

BREW, K., TUMBALE, P. & ACHARYA, K. R. 2010. Family 6 glycosyltransferases in vertebrates and bacteria: inactivation and horizontal gene transfer may enhance mutualism between vertebrates and bacteria. *J Mol Biol,* 285**,** 37121-37127.

BRUNGER, A. T. & ADAMS, P. D. 2002. Molecular Dynamics Applied to X-ray Structure Refinement. *Accounts of Chemical Research,* 35**,** 404-412.

CAMPBELL, J. A., DAVIES, G. J., BULONE, V. & HENRISSAT, B. 1997. A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. *Biochem J,* 326 (Pt 3)**,** 929-39.

CARTE, N., LEGENDRE, F., LEIZE, E., POTIER, N., REEDER, F., CHOTTARD, J. C. & VAN DORSSELAER, A. 2000. Determination by electrospray mass spectrometry of the outersphere association constants of DNA/platinum complexes using 20-mer oligonucleotides and ([Pt(NH(3))(4)](2+), 2Cl(-)) or ([Pt(py)(4)](2+), 2Cl(-)). *Anal Biochem,* 284**,** 77-86.

CHARNOCK, S. J., BERNARD, H. & DAVIES, G. J. 2001. Three-dimensional structures of UDP-sugar glycosyltransferases illuminate the biosynthesis of plant polysaccharides. *Plant Physiology,* 125**,** 527-531.

CHARNOCK, S. J. & DAVIES, G. J. 1999. Cloning, crystallization and preliminary X-ray analysis of a nucleotide-diphospho-sugar transferase spsA from *Bacillus subtilis*. *Acta Crystallographica Section D,* 55**,** 677-678.

CHEN, V. B., ARENDALL, W. B., 3RD, HEADD, J. J., KEEDY, D. A., IMMORMINO, R. M., KAPRAL, G. J., MURRAY, L. W., RICHARDSON, J. S. & RICHARDSON, D. C. 2010. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D,* 66**,** 12-21.

COHEN, S. L. 1996. Domain elucidation by mass spectrometry. *Structure,* 4**,** 1013-6.

COUTINHO, P. M., DELEURY, E., DAVIES, G. J. & HENRISSAT, B. 2003. An evolving hierarchical family classification for glycosyltransferases. *J Mol Biol,* 328**,** 307-317.

DAVIES, G. & HENRISSAT, B. 1995. Structures and mechanisms of glycosyl hydrolases. *Structure,* 3**,** 853-859.

DELANO, W. L. 2010. The PyMOL Molecular Graphics System, Version 1.3. 1.3 ed.: Schrödinger, LLC.

EMSLEY, P. & COWTAN, K. 2004. Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D,* 60**,** 2126-2132.

FRANCO, O. L. & RIGDEN, D. J. 2003. Fold recognition analysis of glycosyltransferase families: further members of structural superfamilies. *Glycobiology,* 13**,** 707-12.

GALILI, U. 2001. The alpha-Gal epitope (Galalpha1-3Galbeta1-4GlcNAc-R) in xenotransplantation. *Biochimie,* 83**,** 557-63.

GALILI, U., CLARK, M. R., SHOHET, S. B., BUEHLER, J. & MACHER, B. A. 1987. Evolutionary relationship between the natural anti-Gal antibody and the Gal alpha 1-3Gal epitope in primates. *Proceedings of the National Academy of Sciences,* 84**,** 1369-1373.

GALILI, U., SHOHET, S. B., KOBRIN, E., STULTS, C. L. & MACHER, B. A. 1988. Man, apes, and Old World monkeys differ from other mammals in the expression of alpha-galactosyl epitopes on nucleated cells. *J Mol Biol,* 263**,** 17755-62.

GARINOT-SCHNEIDER, C., LELLOUCH, A. C. & GEREMIA, R. A. 2000. Identification of essential amino acid residues in the Sinorhizobium meliloti glucosyltransferase ExoM. *J Biol Chem,* 275**,** 31407-13.

GASTEIGER, E., GATTIKER, A., HOOGLAND, C., IVANYI, I., APPEL, R. D. & BAIROCH, A. 2003. ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res,* 31**,** 3784-8.

GASTINEL, L. N., BIGNON, C., MISRA, A. K., HINDSGAUL, O., SHAPER, J. H. & JOZIASSE, D. H. 2001. Bovine [alpha]1,3-galactosyltransferase catalytic domain structure and its relationship with ABO histo-blood group and glycosphingolipid glycosyltransferases. *EMBO J,* 20**,** 638-649.

GASTINEL, L. N., CAMBILLAU, C. & BOURNE, Y. 1999. Crystal structures of the bovine beta4galactosyltransferase catalytic domain and its complex with uridine diphosphogalactose. *EMBO J,* 18**,** 3546-57.

GAVIRA, J. A., HERNANDEZ-HERNANDEZ, M. A., GONZALEZ-RAMIREZ, L. A., BRIGGS, R. A., KOLEK, S. A. & SHAW STEWART, P. D. 2011. Combining Counter-Diffusion and Microseeding to Increase the Success Rate in Protein Crystallization. *Crystal Growth & Design,* 11**,** 2122-2126.

GILLILAND, G. L., TUNG, M., BLAKESLEE, D. M. & LADNER, J. E. 1994. Biological Macromolecule Crystallization Database, Version 3.0: new features, data and the NASA archive for protein crystal growth data. *Acta Crystallographica Section D,* 50**,** 408-13.

GOLOVIN, A., DIMITROPOULOS, D., OLDFIELD, T., RACHEDI, A. & HENRICK, K. 2005. MSDsite: a database search and retrieval system for the analysis and viewing of bound ligands and active sites. *Proteins,* 58**,** 190-9.

GÓMEZ, H., LLUCH, J. M. & MASGRAU, L. 2012. Essential role of glutamate 317 in galactosyl transfer by α3GalT: a computational study. *Carbohydrate Research,* 356**,** 204-208.

GÓMEZ, H., LLUCH, J. M. & MASGRAU, L. 2013. Substrate-Assisted and Nucleophilically Assisted Catalysis in Bovine α1,3-Galactosyltransferase. Mechanistic Implications for Retaining Glycosyltransferases. *Journal of the American Chemical Society,* 135**,** 7053-7063.

GUILLON, P., CLEMENT, M., SEBILLE, V., RIVAIN, J. G., CHOU, C. F., RUVOEN-CLOUET, N. & LE PENDU, J. 2008. Inhibition of the interaction between the SARS-CoV spike protein and its cellular receptor by anti-histo-blood group antibodies. *Glycobiology,* 18**,** 1085-93.

HAKOMORI, S. 1999. Antigen structure and genetic basis of histo-blood groups A, B and O: their changes associated with human cancer. *Biochim Biophys Acta,* 1473**,** 247-66.

HASLAM, D. B. & BAENZIGER, J. U. 1996. Expression cloning of Forssman glycolipid synthetase: a novel member of the histo-blood group ABO gene family. *Proc Natl Acad Sci U S A,* 93**,** 10697-702.

HEADD, J. J., ECHOLS, N., AFONINE, P. V., MORIARTY, N. W., GILDEA, R. J. & ADAMS, P. D. 2014. Flexible torsion-angle noncrystallographic symmetry restraints for improved macromolecular structure refinement. *Acta Crystallographica Section D,* 70**,** 1346-1356.

HEINIG, M. & FRISHMAN, D. 2004. STRIDE: A web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Research,* 32**,** W500-W502.

HEISSIGEROVA, H., BRETON, C., MORAVCOVA, J. & IMBERTY, A. 2003. Molecular modeling of glycosyltransferases involved in the biosynthesis of blood group A, blood group B, Forssman, and iGb(3) antigens and their interaction with substrates. *Glycobiology,* 13**,** 377-386.

HENNET, T. 2002. The galactosyltransferase family. *Cellular and Molecular Life Sciences CMLS,* 59**,** 1081-1095.

HENRISSAT, B., SULZENBACHER, G. & BOURNE, Y. 2008. Glycosyltransferases, glycoside hydrolases: surprise, surprise! *Curr Opin Struct Biol,* 18**,** 527-533.

IGURA, M., MAITA, N., KAMISHIKIRYO, J., YAMADA, M., OBITA, T., MAENAKA, K. & KOHDA, D. 2008. Structure-guided identification of a new catalytic motif of oligosaccharyltransferase. *EMBO J,* 27**,** 234-243.

JAMALUDDIN, H., TUMBALE, P., WITHERS, S. G., ACHARYA, K. R. & BREW, K. 2007. Conformational changes induced by binding UDP-2F-galactose to alpha-1,3 galactosyltransferase- implications for catalysis. *J Mol Biol,* 369**,** 1270-81.

JANCARIK, J. & KIM, S.-H. 1991. Sparse matrix sampling: a screening method for crystallization of proteins. *J Appl Cryst,* 24**,** 409-411.

JINEK, M., CHEN, Y. W., CLAUSEN, H., COHEN, S. M. & CONTI, E. 2006. Structural insights into the Notch-modifying glycosyltransferase Fringe. *Nat Struct Mol Biol,* 13**,** 945-6.

KABSCH, W. & SANDER, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* 22**,** 2577-637.

KAKUDA, S., SHIBA, T., ISHIGURO, M., TAGAWA, H., OKA, S., KAJIHARA, Y., KAWASAKI, T., WAKATSUKI, S. & KATO, R. 2004. Structural basis for acceptor substrate recognition of a human glucuronyltransferase, GlcAT-P, an enzyme critical in the biosynthesis of the carbohydrate epitope HNK-1. *J Biol Chem,* 279**,** 22693-703.

KAMATH, V. P., SETO, N. O., COMPSTON, C. A., HINDSGAUL, O. & PALCIC, M. M. 1999. Synthesis of the acceptor analog alphaFuc(1-->2)alphaGal-O(CH2)7 CH3: a probe for the kinetic mechanism of recombinant human blood group B glycosyltransferase. *Glycoconj J,* 16**,** 599-606.

KANE, J. F. 1995. Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coli. *Curr Opin Biotechnol,* 6**,** 494-500.

KANTARDJIEFF, K. A. & RUPP, B. 2003. Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein–nucleic acid complex crystals. *Protein Science,* 12**,** 1865-1871.

KARLSSON, K. A. 1995. Microbial recognition of target-cell glycoconjugates. *Curr Opin Struct Biol,* 5**,** 622-35.

KENDREW, J. C., BODO, G., DINTZIS, H. M., PARRISH, R. G., WYCKOFF, H. & PHILLIPS, D. C. 1958. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature,* 181**,** 662-6.

KEUSCH, J. J., MANZELLA, S. M., NYAME, K. A., CUMMINGS, R. D. & BAENZIGER, J. U. 2000a. Cloning of Gb3 synthase, the key enzyme in globo-series glycosphingolipid synthesis, predicts a family of alpha 1, 4-glycosyltransferases conserved in plants, insects, and mammals. *J Biol Chem,* 275**,** 25315-21.

KEUSCH, J. J., MANZELLA, S. M., NYAME, K. A., CUMMINGS, R. D. & BAENZIGER, J. U. 2000b. Expression Cloning of a New Member of the ABO Blood Group Glycosyltransferases, iGb3 Synthase, That Directs the Synthesis of Isoglobo-glycosphingolipids. *J Biol Chem,* 275**,** 25308-25314.

KLEYWEGT, G. J. 1996. Use of non-crystallographic symmetry in protein structure refinement. *Acta crystallographica Section D, Biological crystallography,* 52**,** 842-57.

KOIKE, C., FUNG, J. J., GELLER, D. A., KANNAGI, R., LIBERT, T., LUPPI, P., NAKASHIMA, I., PROFOZICH, J., RUDERT, W., SHARMA, S. B., STARZL, T. E. & TRUCCO, M. 2002. Molecular basis of evolutionary loss of the alpha 1,3-galactosyltransferase gene in higher primates. *J Biol Chem,* 277**,** 10114-20.

KRISSINEL, E. & HENRICK, K. 2007. Inference of macromolecular assemblies from crystalline state. *J Mol Biol,* 372**,** 774-797.

LAIRSON, L. L., CHIU, C. P. C., LY, H. D., HE, S., WAKARCHUK, W. W., STRYNADKA, N. C. J. & WITHERS, S. G. 2004. Intermediate trapping on a mutant retaining α-galactosyltransferase identifies an unexpected aspartate residue. *J Mol Biol,* 279**,** 28339-28344.

LAIRSON, L. L., HENRISSAT, B., DAVIES, G. J. & WITHERS, S. G. 2008. Glycosyltransferases: Structures, functions, and mechanisms. *Annual Review of Biochemistry,* 77**,** 521-555.

LANTERI, M., GIORDANENGO, V., VIDAL, F., GAUDRAY, P. & LEFEBVRE, J. C. 2002. A complete alpha1,3-galactosyltransferase gene is present in the human genome and partially transcribed. *Glycobiology,* 12**,** 785-92.

LARKIN, M. A., BLACKSHIELDS, G., BROWN, N. P., CHENNA, R., MCGETTIGAN, P. A., MCWILLIAM, H., VALENTIN, F., WALLACE, I. M., WILM, A., LOPEZ, R., THOMPSON, J. D., GIBSON, T. J. & HIGGINS, D. G. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics,* 23**,** 2947-2948.

LASKOWSKI, R. A., MACARTHUR, M. W., MOSS, D. S. & THORNTON, J. M. 1993. {PROCHECK}: a program to check the stereochemical quality of protein structures. *J Appl Cryst,* 26**,** 283-291.

LEE, H. J., BARRY, C. H., BORISOVA, S. N., SETO, N. O., ZHENG, R. B., BLANCHER, A., EVANS, S. V. & PALCIC, M. M. 2005. Structural basis for the inactivity of human blood group O2 glycosyltransferase. *J Biol Chem,* 280**,** 525-9.

LEE, S. S., HONG, S. Y., ERREY, J. C., IZUMI, A., DAVIES, G. J. & DAVIS, B. G. 2011. Mechanistic evidence for a front-side, SNi-type reaction in a retaining glycosyltransferase. *Nat Chem Biol,* 7**,** 631-638.

LIU, J. & MUSHEGIAN, A. 2003. Three monophyletic superfamilies account for the majority of the known glycosyltransferases. *Protein Sci,* 12**,** 1418-31.

LOMBARD, V., GOLACONDA RAMULU, H., DRULA, E., COUTINHO, P. M. & HENRISSAT, B. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research,* 42**,** D490-D495.

MATTHEWS, B. W. 1968. Solvent content of protein crystals. *J Mol Biol,* 33**,** 491-7.

MCCOY, A. J., GROSSE-KUNSTLEVE, R. W., ADAMS, P. D., WINN, M. D., STORONI, L. C. & READ, R. J. 2007. Phaser crystallographic software. *J Appl Cryst,* 40**,** 658-674.

MCPHERSON, A. 1999. *Crystallisation of biological macromolecules,* Cold Spring Harbor Press, New York.

MCPHERSON, A. & GAVIRA, J. A. 2014. Introduction to protein crystallization. *Acta Crystallographica Section F,* 70**,** 2-20.

MCPHERSON, A., KUZNETSOV, Y. G., MALKIN, A. & PLOMP, M. 2003. Macromolecular crystal growth as revealed by atomic force microscopy. *Journal of Structural Biology,* 142**,** 32-46.

MONEGAL, A. & PLANAS, A. 2006. Chemical rescue of alpha3-galactosyltransferase. Implications in the mechanism of retaining glycosyltransferases. *J Am Chem Soc,* 128**,** 16030-1.

NEIL, S. J., MCKNIGHT, A., GUSTAFSSON, K. & WEISS, R. A. 2005. HIV-1 incorporates ABO histo-blood group antigens that sensitize virions to complement-mediated inactivation. *Blood,* 105**,** 4693-9.

OHTSUBO, K., IMAJO, S., ISHIGURO, M., NAKATANI, T., OKA, S. & KAWASAKI, T. 2000. Studies on the structure-function relationship of the HNK-1 associated glucuronyltransferase, GlcAT-P, by computer modeling and site-directed mutagenesis. *J Biochem,* 128**,** 283-91.

OTWINOWSKI, Z. & MINOR, W. 1997. [20] Processing of X-ray diffraction data collected in oscillation mode. *In:* CHARLES W. CARTER, JR. (ed.) *Methods in Enzymology.* Academic Press. pp 307-326.

PAK, J. E., ARNOUX, P., ZHOU, S., SIVARAJAH, P., SATKUNARAJAH, M., XING, X. & RINI, J. M. 2006. X-ray crystal structure of leukocyte type core 2

beta1,6-N-acetylglucosaminyltransferase. Evidence for a convergence of metal ion-independent glycosyltransferase mechanism. *J Biol Chem,* 281**,** 26693-701.

PAK, J. E., SATKUNARAJAH, M., SEETHARAMAN, J. & RINI, J. M. 2011. Structural and mechanistic characterization of leukocyte-type core 2 β1,6-N-acetylglucosaminyltransferase: A metal-ion-independent GT-A glycosyltransferase. *J Mol Biol,* 414**,** 798-811.

PATENAUDE, S. I., SETO, N. O., BORISOVA, S. N., SZPACENKO, A., MARCUS, S. L., PALCIC, M. M. & EVANS, S. V. 2002. The structural basis for specificity in human ABO(H) blood group biosynthesis. *Nat Struct Biol,* 9**,** 685-90.

PREECE, A. F., STRAHAN, K. M., DEVITT, J., YAMAMOTO, F. & GUSTAFSSON, K. 2002. Expression of ABO or related antigenic carbohydrates on viral envelopes leads to neutralization in the presence of serum containing specific natural antibodies and complement. *Blood,* 99**,** 2477-82.

QASBA, P. K., RAMAKRISHNAN, B. & BOEGGEMAN, E. 2005. Substrate-induced conformational changes in glycosyltransferases. *Trends Biochem Sci,* 30**,** 53-62.

RHODES, G. 2006. *Crystallography made crystal clear : a guide for users of macromolecular models,* Amsterdam ; Boston, Elsevier/Academic Press.

RINI, J., ESKO, J. & VARKI, A. 2009. Glycosyltransferases and Glycan-processing Enzymes. *In:* VARKI, A., CUMMINGS, R. D., ESKO, J. D., FREEZE, H. H., STANLEY, P., BERTOZZI, C. R., HART, G. W. & ETZLER, M. E. (eds.) *Essentials of Glycobiology.* 2nd ed. Cold Spring Harbor (NY).

RUPP, B. 2010. *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology,* Garland Science.

RUSSO KRAUSS, I., MERLINO, A., VERGARA, A. & SICA, F. 2013. An overview of biological macromolecule crystallization. *Int J Mol Sci,* 14**,** 11643-91.

SAITOH, S., NODA, S., AIBA, Y., TAKAGI, A., SAKAMOTO, M., BENNO, Y. & KOGA, Y. 2002. Bacteroides ovatus as the predominant commensal intestinal

microbe causing a systemic antibody response in inflammatory bowel disease. *Clin Diagn Lab Immunol,* 9**,** 54-9.

SEYMOUR, R. M., ALLAN, M. J., POMIANKOWSKI, A. & GUSTAFSSON, K. 2004. Evolution of the human ABO polymorphism by two complementary selective pressures. *Proc Biol Sci,* 271**,** 1065-72.

SLEDZ, P., ZHENG, H., MURZYN, K., CHRUSZCZ, M., ZIMMERMAN, M. D., CHORDIA, M. D., JOACHIMIAK, A. & MINOR, W. 2010. New surface contacts formed upon reductive lysine methylation: Improving the probability of protein crystallization. *Protein Science,* 19**,** 1395-1404.

SODOYER, R. 2004. Expression systems for the production of recombinant pharmaceuticals. *BioDrugs,* 18**,** 51-62.

SOYA, N., FANG, Y., PALCIC, M. M. & KLASSEN, J. S. 2011. Trapping and characterization of covalent intermediates of mutant retaining glycosyltransferases. *Glycobiology,* 21**,** 547-552.

SOYA, N., SHOEMAKER, G. K., PALCIC, M. M. & KLASSEN, J. S. 2009. Comparative study of substrate and product binding to the human ABO(H) blood group glycosyltransferases. *Glycobiology,* 19**,** 1224-34.

STAMPS, R., SOKOL, R. J., LEACH, M., HERRON, R. & SMITH, G. A. 1987. A new variant of blood group A. *Transfusion,* 27**,** 315-318.

SUN, H. Y., LIN, S. W., KO, T. P., PAN, J. F., LIU, C. L., LIN, C. N., WANG, A. H. & LIN, C. H. 2007. Structure and mechanism of Helicobacter pylori fucosyltransferase. A basis for lipopolysaccharide variation and inhibitor design. *J Biol Chem,* 282**,** 9973-82.

SVENSSON, L., HULT, A. K., STAMPS, R., ÅNGSTRÖM, J., TENEBERG, S., STORRY, J. R., JØRGENSEN, R., RYDBERG, L., HENRY, S. M. & OLSSON, M. L. 2013. Forssman expression on human erythrocytes: biochemical and genetic evidence of a new histo-blood group system. *Blood,* 121**,** 1459-1468.

TERPE, K. 2006. Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems. *Appl Microbiol Biotechnol,* 72**,** 211-22.

THIYAGARAJAN, N., PHAM, T. T., STINSON, B., SUNDRIYAL, A., TUMBALE, P., LIZOTTE-WANIEWSKI, M., BREW, K. & ACHARYA, K. R. 2012. Structure of a metal-independent bacterial glycosyltransferase that catalyzes the synthesis of histo-blood group A antigen. *Sci Rep,* 2**,** 940.

TU, L. & BANFIELD, D. K. 2010. Localization of Golgi-resident glycosyltransferases. *Cell Mol Life Sci,* 67**,** 29-41.

TUMBALE, P. & BREW, K. 2009. Characterization of a metal-independent CAZy family 6 glycosyltransferase from Bacteroides ovatus. *J. Biol. Chem.,* 284**,** 25126-25134.

TUMBALE, P., JAMALUDDIN, H., THIYAGARAJAN, N., BREW, K. & ACHARYA, K. R. 2008. Structural basis of UDP-galactose binding by alpha-1,3-galactosyltransferase (alpha3GT): role of negative charge on aspartic acid 316 in structure and activity. *Biochemistry,* 47**,** 8711-8.

UNLIGIL, U. M., ZHOU, S., YUWARAJ, S., SARKAR, M., SCHACHTER, H. & RINI, J. M. 2000. X-ray crystal structure of rabbit N-acetylglucosaminyltransferase I: catalytic mechanism and a new protein superfamily. *EMBO J,* 19**,** 5269-80.

VAGIN, A. & TEPLYAKOV, A. 1997. MOLREP: an Automated Program for Molecular Replacement. *J Appl Cryst,* 30**,** 1022-1025.

VELANKAR, S., MCNEIL, P., MITTARD-RUNTE, V., SUAREZ, A., BARRELL, D., APWEILER, R. & HENRICK, K. 2005. E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res,* 33**,** D262-5.

VRIELINK, A., RUGER, W., DRIESSEN, H. P. & FREEMONT, P. S. 1994. Crystal structure of the DNA modifying enzyme beta-glucosyltransferase in the presence and absence of the substrate uridine diphosphoglucose. *EMBO J,* 13**,** 3413-22.

WALLACE, A. C., LASKOWSKI, R. A. & THORNTON, J. M. 1995. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng,* 8**,** 127-34.

WINN, M. D., BALLARD, C. C., COWTAN, K. D., DODSON, E. J., EMSLEY, P., EVANS, P. R., KEEGAN, R. M., KRISSINEL, E. B., LESLIE, A. G. W., MCCOY, A., MCNICHOLAS, S. J., MURSHUDOV, G. N., PANNU, N. S., POTTERTON, E. A., POWELL, H. R., READ, R. J., VAGIN, A. & WILSON, K. S. 2011. Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D,* 67**,** 235-242.

WINTER, G. 2010. xia2: an expert system for macromolecular crystallography data reduction. *J Appl Cryst,* 43**,** 186-190.

WLODAWER, A., MINOR, W., DAUTER, Z. & JASKOLSKI, M. 2008. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS J,* 275**,** 1-21.

YAMAMOTO, F.-I., MCNEILL, P. D. & HAKOMORI, S.-I. 1995. Genomic organization of human histo-blood group ABO genes. *Glycobiology,* 5**,** 51-58.

YAMAMOTO, F. 2000. Molecular genetics of ABO. *Vox Sang,* 78 Suppl 2**,** 91-103.

YAMAMOTO, F., CLAUSEN, H., WHITE, T., MARKEN, J. & HAKOMORI, S. 1990. Molecular genetic basis of the histo-blood group ABO system. *Nature,* 345**,** 229-33.

ZHANG, Y., DESHPANDE, A., XIE, Z., NATESH, R., ACHARYA, K. R. & BREW, K. 2004. Roles of active site tryptophans in substrate binding and catalysis by alpha-1,3 galactosyltransferase. *Glycobiology,* 14**,** 1295-302.

ZHANG, Y., SWAMINATHAN, G. J., DESHPANDE, A., BOIX, E., NATESH, R., XIE, Z., ACHARYA, K. R. & BREW, K. 2003. Roles of individual enzyme-substrate interactions by alpha-1,3-galactosyltransferase in catalysis and specificity. *Biochemistry,* 42**,** 13512-21.

ZHANG, Y., WANG, P. G. & BREW, K. 2001. Specificity and mechanism of metal Ion activation in UDP-galactose:β-galactoside-α-1,3-galactosyltransferase. *J Mol Biol,* 276**,** 11567-11574.

APPENDIX

Appendix

Table A1. Interactions between BoGT6a E192Q with intact UDP-GalNAc

Ligand	Distance (Å)		Residues	Chain
	Min	Max		
[O6']	2.92		TRP 189[O]	P
	3.37		GLN 192[OE1]	P
[N2']	3.13	3.73	ASN 95[OD1]	G, H, O
[N3]	2.51	3.25	THR 10[OG1]	E, F, G, H, O
	2.32	3.39	THR 10[O]	E, F, G, H, P
[O1B]	3.69		LYS 231[NZ]	F
	3.90		ASN 95[ND2]	H
[O2B]	3.20	3.49	ARG 243[NH2]	E, G
[O1']	3.70		ASN 95[ND2]	E
[O6']	3.71		HIS 190[N]	E
	3.70		GLN 192[NE2]	G
[O4']	3.44	3.86	GLN 192[N]	E, F, H, O
	2.97		GLN 192[NE2]	G
[O3']	2.83	3.62	GLY 157[N]	E, F, G, H, O, P
	2.68	3.59	ARG 73[NH2]	E, F, G, H, O, P
[O7']	3.28	3.57	ASN 95[ND2]	E, F, P
[O3A]	2.30	2.57	LYS 231[NZ]	E, G, O
	3.33	3.45	ARG 243[NH2]	F, H, P

Ligand	Distance (Å)		Residues	Chain
	Min	Max		
[O1A]	2.55	2.94	TYR 13[OH]	E, G
	2.54	2.80	LYS 231[NZ]	F, H, O
	3.28	3.61	ASN 95[ND2]	F, H, P
	3.10		ARG 243[NH2]	G
[O2A]	3.14	3.31	ARG 243[NH2]	F, H, P
	2.64	2.93	TYR 13[OH]	F, H, P
[O3B]	2.65	3.68	ALA 96[N]	E, F, G, H, O, P
	2.98		ASN 95[ND2]	O
[O2]	3.23		THR 10[OG1]	E
	2.49	3.61	THR 10[N]	E, F, G, H, O, P
[O5B]	3.81		ASN 95[ND2]	H
[O4]	2.75		ASN 69[ND2]	H

Table A2. Interactions between BoGT6a E192Q with UDP and GalNAc

Ligand	Distance		Residues	Chain
	Min	Max		
UDP				
[N3]	3.54	3.85	ASN 69[OD1]	A, K
	2.40	3.30	THR 10[OG1]	A, B, C, D, I, J, M
	2.62	3.40	THR 10[O]	A, B, D, I, J, K, L, M

Ligand	Distance		Residues	Chain
	Min	Max		
[O1B]	2.23	3.42	TYR 13[OH]	A, J, L
	2.60	2.75	LYS 231[NZ]	M, N
[O2B]	2.50	2.74	LYS 231[NZ]	A, I
	2.67	3.00	TYR 13[OH]	C, I
[O3B]	2.97		TYR 13[OH]	B
	3.29		TRP 66[NE1]	C
	2.47		LYS 231[NZ]	J
[O1A]	2.97	3.77	TYR 13[OH]	A, K, N
	2.21	2.38	LYS 231[NZ]	D, M, N
[O2A]	2.33	2.44	LYS 231[NZ]	B, I, L
	3.68	3.85	ARG 243[NH2]	M, N
[O3A]	3.24		TYR 13[OH]	B
	3.22		LYS 231[NZ]	C
[O3']	2.76	3.07	ALA 96[N]	A, B, C, D, I, J, L, M, N
[O2']	3.45	3.77	ALA 96[N]	A, C, L
	3.85		CYS 9[N]	M
[O2]	2.83	3.41	THR 10[N]	A, B, C, D, I, J, K, L, M, N
	2.55	3.36	THR 10[OG1]	B, I, J, K
[O5']	3.48		LYS 231[NZ]	C
[O4]	2.55	3.37	ASN 69[ND2]	C, L, M

Ligand	Distance		Residues	Chain
	Min	Max		
[O4']	3.69		THR 70[OG1]	M
α - GalNAc				
[O1]	2.50	2.60	HIS 122[NE2]	A, K
[O]	3.62		GLN 192[NE2]	A
[O6]	3.18	3.76	THR 134[OG1]	B, D
	2.58		GLN 192[NE2]	D
	2.80		TRP 189[NE1]	I
[O4]	3.01		GLN 192[NE2]	B
	2.62		HIS 122[NE2]	I
[O3]	2.87		GLN 192[NE2]	C
β - GalNAc				
[C1]	1.30	1.31	GLN 192[NE2]	M, N
[O4]	3.38	3.69	ARG 73[NH2]	M, N
[O3]	2.60	2.81	ASN 95[ND2]	M, N
[O6]	3.59		ASP 191[N]	N
[O7]	2.46		ASN 95[ND2]	N